QUALITATIVE COMPARISON OF POSTAL ADDRESS ELEMENT LISTS

The purpose of this paper is to compare six different lists of postal address elements to determine the underlying methodological differences among them. Differences in the elements included in the lists themselves often flow out of these methodological differences, so this is a helpful way to gain understanding. The actual differences in the elements lists themselves will be cited where they illustrate the argument.

After discussing the methodological differences, we will present a position on several issues pertaining to how the UPU address standardization effort should proceed both with respect to the task of developing an international postal addressing standard and in order to promote harmonization with related standards that are likely to be influential in electronic commerce.

The strengths and weaknesses of the various standardization proposals from the point of view of the author are also listed in the hope that this will further encourage discussion and lead to stronger standards in the long run.

The six candidates are the CEN proposal dated February 11, 2002 on Components of Postal Addresses (CEN), the current ECCMA list of international address element codes (ECCMA), a relevant subset of the USPS Address Management System data base fields (AMS) as reported by Mabel Grein of USPS, the IDEAlliance Address Data Interchange Specification version 01-1 (ADIS), the OASIS CIQ TC Extensible Name and Address Language draft version 2.0 (xNAL), and the HR-XML consortium PostalAddress 1.2 and PersonName 1.2 (HR-XML).

The dimensions we will consider are hierarchical depth, granularity, backward compatibility, technology utilized, postal specificity, significance of sequential order, cardinality, breadth of international capabilities, and suitability for use with templates. There could certainly be other dimensions considered, but it is necessary to start somewhere. It is also recognized that what may seem a virtue in one context may be a drawback in another context, so that locating an address element list on a dimension does not in itself constitute a positive or negative evaluation. Finally, all the judgments made are subjective and should not be viewed as definitive.

With those demurrals, we enter the fray by evaluating the hierarchical depth of each listing. This corresponds roughly to the level of nesting of the elements, although it is possible to try for both hierarchical depth and a capability for flattened direct access to lower levels. That having been said, the six candidates appear to rank in the following order of higher to lower hierarchical depth: xNAL, CEN, HR-XML, ADIS, ECCMA and AMS. Greater depth provides a major advantage in terms of clarity, as long as the categories are carefully thought out. At the same time it adds complexity, and efforts to circumvent the hierarchy run the risk of having different applications of the specification at different levels of depth, which can raise compatibility issues.

By granularity, we mean the level of detail and the atomicity of the lower level elements. The lists that appear to have high granularity include ECCMA, xNAL, and ADIS, while the others, namely HR-XML, AMS, and CEN, have medium granularity, in the sense that some of the lowest level elements are not atomic.

This must be considered in the context of which lists try for backward compatibility with legacy applications by allowing for aggregations of elements in lieu of exclusive reliance on parsed elements. The lists that allow backward compatibility are trying to be reasonable and practical about the issues of transition from a world that is address block based or at best line based to an element based environment. At the same time this methodology risks compromising the benefits of an element based approach, and could lead to incompatible applications while dressing up the endemic shortcomings of address management technology in new technological clothes. The lists that support backward compatibility include ADIS, xNAL and HR-XML, while AMS is a hybrid, and those that are more purely element based include ECCMA and CEN.

In terms of technology utilized, HR-XML and xNAL provide XML schemas and DTDs, ADIS has a DTD and a database format, ECCMA is a code list, CEN provides a natural language description at this time, and AMS is a database. With respect to the schemas and DTDs, both HR-XML and xNAL aim for compatibility between the two representations. Our viewpoint is that this gives up much of the power of XML schemas. Several problems discussed below relating to the usability of address element list specifications for postal purposes, such as cardinality and sequential occurrences, could be approached more productively by allowing the XML schemas to diverge from the DTDs. Nevertheless, there are still some advantages to DTDs in terms of their ability to communicate the semantics of complex data structures, while the most powerful schemas may be almost completely opaque to the inquiring reader.

The attribute of postal specificity relates to the intended use of the address lists. Those lists that are specific to the postal realm may have the merit of being more directly applicable to a narrow range of uses such as preparing mailings. Those that are more generic have the merit of being useful across a wide variety of applications beyond the postal realm, such as CRM, data quality, or non-postal forms of distribution. The lists with postal specificity are likely to branch out in the direction of support for mail production and obtaining postal discounts. The lists with a more general approach to addresses are likely to branch out into areas such as marketing, geocoding, and customer information. Of those considered, CEN, ECCMA, ADIS and AMS are intended to have postal specificity, while HR-XML and xNAL are more “application independent”.

Some of the lists rely on the sequential order in which the data instances are first acquired to populate the elements in such a way that they can be reconstituted and still make sense. Others try for a limit the dependence on the original sequential order, usually by insisting on or at least aiming to maintain a functional definition of each element. Examples can be found in both name components such as surname and address components such as locality. Among those explicitly preserving the sequential order are ADIS and HR-XML. All of the others, including xNAL, CEN, AMS and ECCMA, use this methodology in some situations, or at least allow the user to do so. This widespread practice shows a difficulty that perhaps all manifestations of address element technology must overcome. While parsing data into its distinct components, the meaning must still be preserved. Sometimes the original sequential order of the components is essential to the meaning, while at other times it can be arbitrary or even erroneous.

This issue is related to cardinality, or the number of instances of each element that are allowed. With a cardinality of greater than one, there is an implicit issue of sequential order. For example, with two addressees, is the first one listed primary or is the order arbitrary? CEN has yet to fully commit itself on this issue, and ECCMA lacks a workable method to cope with multiple street address lines or multiple secondary unit identifiers and designators. This is partly because the ECCMA codes reflect the underlying AMS database which is intentionally limited in cardinality. OASIS xNAL and HR-XML, originally XML DTDs, swing back and forth between the alternatives of a cardinality of 0 to 1 and the other extreme, which allows infinite occurrences. ADIS in several cases establishes a maximum number of instances of two, and in one case of four, by the method of repetition. This may have some practical usefulness, but inevitably risks counterexamples.

Whatever is said about breadth of international coverage might be disputed. Perhaps it is a good sign that many of the lists have begun to borrow elements from each other, becoming more comprehensive in the process. It seems to the author that xNAL in its most recent version has the greatest breadth, followed by CEN, with HR-XML having given some thought to the matter, and AMS, ECCMA, and ADIS more limited, showing their U. S. and North American origins. Of course, it is not fair to complain about AMS, the product of the USPS, in this respect. ECCMA has just received a list of Japanese candidate elements, some of which may be redundant. As for ADIS, if it evolves along the lines set down in the Postal Address Template Description Language (PATDL) paper, it will acquire the ability to use external element lists.

Since the UPU address standardization project calls for the use of templates to determine the sequencing and formatting of elements, a further consideration for the lists is how well suited they are to be used within templates. This again may be a matter of judgment, but ADIS and ECCMA were certainly designed with templates in mind, most notably in using codes rather than only natural language tags to denote the elements. But ECCMA, as noted above, lacks adequate support for cardinality in areas such as multiple street address lines and secondary unit identifiers and designators. OASIS xNAL now allows for the use of external code sets, but may be hampered by the potential presence of data at multiple hierarchical levels. Lacking any support for codes, the CEN and HR-XML lists are nonetheless capable of being referenced in templates, though with some difficulty. The CEN list needs support for multiple sequential occurrences of its elements, and in general for multiple occurrences of its elements, to represent the full range of data that can occur within postal addresses. The unlimited cardinality at certain places in the xNAL and HR-XML specifications can pose the opposite problem for the template designer and in the rendition process. Addresses represented in either of these specifications may overflow limits on cardinality or on multiple sequential occurrences of elements within templates. As for AMS, it is intended as a target for matching logic rather than as a list of codes or tags.

How does each of the six lists specifically handle names? For AMS the answer is simple. For legal reasons there are no names in AMS. The USPS has names in its National Change of Address (NCOA) database, and some other related files, because that data is related to the legally sanctioned purpose of tracking move updates. But the NCOA files are separate from AMS.

ECCMA allows for preceding honorifics such as “Mr” or “Dr”, and then sequentially designates first name, first middle name, second middle name, and last name, followed by generation, such as “Jr” or “III”, and suffix such as “PhD”. There is a field for preferred name or nickname, in other words “Bob” instead of “Robert”. ECCMA allots fields for a second addressee, but no nickname for that individual. In related fields, mailee and professional title are designated. Consistent with the ECCMA methodology, no unparsed name fields are provided.

ADIS is similar to ECCMA, but also allows for a last name prefix, such as “von”, a nickname for both of two addressees, and an alternate postal recipient field that can store the notation “Or Current Resident” or similar forms. Following a hybrid approach rather than exclusively element based, ADIS allows for unparsed name data, though the ADIS DTD is set up to deny the opportunity for the data to be stored in both a parsed and an unparsed manner. It also has a field for mailee and professional title.

CEN has similar fields to those considered above using different field names, such as form of address instead of pre-honorific, name qualifier instead of suffix, and qualification instead of post-honorific. For the name itself, CEN has given name, surname prefix, and surname root. Middle name and preferred name can be stored in the given name field. This set of definitions may be too culturally specific to apply worldwide.

A particularly strong feature of the CEN element list is the careful delineation of roles in a mailing situation. Rather than just allowing for a mailee, CEN defines an addressee, a mail originator, a mail recipient, a mail submitter, a mailee, a mailer and a payer. These are very precise distinctions, though in the further development of a model, for a large mailing file, some of these roles, such as mailer, payer and mail submitter, need not be repeated each time.

HR-XML allows for a formatted name and a legal name separately from the parsed name fields. There is a given name with unlimited cardinality, a preferred name, a middle name with unlimited cardinality, and a family name with unlimited cardinality. Attributes can optionally be used to designate certain fields as primary to assist in the formatting tasks.

OASIS xNAL supports a multiplicity of formats for names. Where appropriate, the intermediate level fields have unlimited cardinality. There is an unparsed name capability. There can be the name of a person and separately the name of a joint person, which constitutes a type of subordinate recipient. The fields within the person name container are defined sequentially. The prefix to the last name can be broken out separately. There are additional fields such as former name, alias, and the name by which a person is known. This is appropriate for a generic rather than specifically postal approach. It could be used, for example, by a company with long-term client relationships, or by a government agency. A potential difficulty with the new version is that the unparsed name line can occur in several different places, and not just because of the alternate name capability. For example, the name of the primary addressee could appear outside of the “PersonName” element, or inside that element. This gives the user flexibility but makes it hard on an implementer who wants to support the full specification, or on a template designer.

How does each of the lists handle address data? AMS, because of its special purpose, has some database fields that contain range data, such as for the primary number. With respect to street names, there are separate fields for pre-directional, street name, post-directional, and suffix, because these are elements that are very common in US addresses. There is no preceding street type, so “Avenida Colon” would be found in the street name field. Post office box numbers and non-postal box numbers are stored separately, but military addresses do not have their own dedicated fields. There is a single secondary type, such as "Apt" for apartment, with a range of apartment numbers possible. This means that "Building 7 Unit 3" can cause some difficulties. There is a single field that is apparently used for building name, firm name, and the names of institutions such as schools. This saves space in the database, though there are times when more than one instance would be desirable.

AMS also supports the urbanization field for Puerto Rico, the city name, state abbreviation, and naturally, the ZIP code in its various forms.

ECCMA generally works on the assumption that there is one data record per delivery point, though there is at least one set of range fields that is supported. The street address, with a cardinality of one, consists of primary number, pre-directional, street name, street name suffix, post-directional, and one secondary identifier and designator. Rural routes and post office boxes have their own fields. Unparsed street address data is not supported. Business information includes firm name, organization name and type, department, division, building name and type, and mail stop. ECCMA supports most of the CEN street address elements where these differ from what is mentioned above.

For the last lines, ECCMA supports the AMS fields, and the CEN fields such as dependent locality, international postal code, and receiving country. The bar code check digits for the various ZIP code formats are supported.

ADIS supports two street address lines and four secondary identifiers and designators for each of the street address lines. This was done at the request of mailing industry experts that find a need for that many fields in the address data streams exchanged by businesses. There is a preceding street type field, useful for special situations in the US and found commonly in various countries. The rural route type and designator are separated into two fields. Two unparsed address lines are supported, though the DTD prevents both parsed and unparsed data from being carried at the same time. The business information is parallel to that carried in ECCMA. Most of the CEN street name fields are included in ADIS, where not already mentioned.

In the last lines, ADIS supports the ECCMA fields, includes an unparsed last line, and allows documentation of the sending country. The PLANET code, used for mail tracking, is also supported in ADIS.

The CEN address elements include street number, thoroughfare name, type and qualifier, along with thoroughfare access data. Also included are building, building type, wing, floor, and door, extension designation, and supplemental delivery data. Within an organization there are elements for name, unit, function and legal status. The locality construct includes region, town, district, and distribution area indicator. Country and postcode are supported, as well as the defining authority responsible for delivery.

The HR-XML address elements include an unparsed address line, and a combination of street name, building number, unit, and postal box. The building number field is where the primary house number can be stored. There is an admonition not to use both parsed and unparsed data for the same address. An address type field is defined, and provision is made for a country code.

In addition to the street address, multiple occurrences of region and a municipality and postcode are provided for.

The OASIS xNAL address structure is deeply nested. A key construct is thoroughfare which can contain a number or range, prefix, suffix, pre-directional, leading type, street name, trailing type, post-directional, dependent thoroughfare, dependent locality, premise, firm and postal code. It can also contain an unparsed address line.

Above the thoroughfare construct are “AddressDetails”, which includes an entire unparsed address, an area that can contain identified but unparsed address lines, and containers for country, administrative area, locality, and thoroughfare. There is also support for address types, address codes, and postal service related data to be stored within “AddressDetails”. The postal service related data supports some of the information needed for mailing production, such as barcodes. The address codes could include codes from other element lists.

Administrative area includes an unparsed line capability as well as the name of the area, a container for sub-administrative area, and optional locality, post office, and postal code. Sub-administrative area includes an unparsed line capability as well as the name of the area, and optional locality, post office, and postal code.

Locality includes an unparsed line capability as well as the locality name, options for a post box, large mail user identifier, post office, or rural route, and further options for thoroughfare, premise, dependent locality, and postal code. The reader should note that several of the containers for fields are reused at different levels of the hierarchy where appropriate.

Other elements that contain sub-elements include country, post office, postal code, post box, premise, sub-premise, dependent locality, dependent thoroughfare, firm, postal route, post town, large mail user, department, mail stop, and ranges for thoroughfare numbers and premise numbers. While this version of the xNAL specification is in draft status and may experience some modifications, it is probable that it can store almost any address. The elements that need a cardinality of more than one are generally provided with infinite cardinality. On the other hand, different users might store the same address in a myriad of ways, which raises some issues of interoperability.

In summary, each of the six approaches has its characteristic strengths and weaknesses. The AMS database has been the foundation for most of the USPS address management capabilities, which have been a major asset to the postal service that handles 40% of the world’s mail. Nonetheless, it could be improved by adding certain extensions, and its way of handling certain situations such as military addresses and Puerto Rico addresses is less than ideal.

The strongest feature of the ECCMA codes is their simplicity and atomicity, which provides for ease of implementation. They can be implemented in XML, EDI or database technology. The concept can clearly be extended to cover other geographic areas that are underrepresented. In the future, they may become somewhat eclectic as users add on to what is already there to meet disparate needs.

ADIS has strengths in the area of business mail production and its support for templates and rendition instructions. It needs to improve in the area of internationalization and precise definition of its elements, and it lacks an XML schema. By evolving in the direction of the PATDL paper, ADIS can provide an early implementation of the developing UPU standard.

The CEN element list is strongest in its conceptual foundation, which justifies its use as the basis for the UPU definition of address elements. The early work on "printing rules" from the CEN also shows considerable promise. The CEN element list needs to evolve in the direction of a formal model, and it should be opened up to support lower level elements from other parts of the world within its framework of segments and constructs.

The HR-XML standards are easy to understand and would probably be relatively simple to implement. They are better suited for the general business environment, and less well suited for postal specialists or for advanced address hygiene applications.

The OASIS xNAL specification is the most complete XML definition of international name and address elements. It is likely to be a strong candidate for implementation in a variety of electronic commerce applications. By design, it is not intended to support specific postal applications such as mail production and final presentation of the address. With its deep hierarchical structure, a full implementation of the specification is likely to be complex.

What course of action is recommended for the POST*Code group, based on progress so far and the results of the comparison of alternative lists of address elements? In my view, there is not room for six contenders in the area of address element technology, but there is room for at least two. One of these can be specific to postal applications. It can aim at maintaining postal addresses in the form preferred by the postal services, and at facilitating the value chain that runs from address acquisition through address management to mailing production. As the UPU resolutions state, it should encompass elements, templates, and rendition instructions. In the view of the author, by combining the best features of CEN, ECCMA, AMS and ADIS, and using insights from evaluating HR-XML and xNAL, such a standard can now move forward.

There is also a need for a generic address standard suitable for address management, marketing, CRM, and other functions involving long-term client relationships. The contenders here are clearly HR-XML and xNAL. Both of these are intended for use within a wide variety of business processes, and neither is intended to describe final presentation on a mail piece. As the more comprehensive of the two, xNAL is well positioned in this area, providing it does not prove overly complex in implementation, and can avert the risk of incompatibility among its own applications. The unlimited cardinality that is found at key places in HR-XML and xNAL is a disadvantage for postal address template design and in the rendition process, but provides flexibility and extensibility for "application independent" purposes.

In any event, there is a need for interoperability among all the address element technology efforts. It would not be easy to resolve the differences among these efforts at the level of methodological premises. It also would probably be difficult to gain consensus at the level of segments and constructs. The opportunity for interoperability may be most promising in the area of lower level elements. One recommendation of this paper is that the ECCMA code list be further extended to encompass as many of the lower level elements as possible from the other lists. All the elements in the eventual UPU standard should have ECCMA codes, though it is not necessary that all the ECCMA elements become part of the UPU standard element list. Since the ECCMA code list is relatively lacking in a high level conceptual structure with segments and constructs, this may be the easiest terrain on which to pursue interoperability. But this is not simply a matter of adding elements. It will also be necessary to address the issues of cardinality and multiple sequential occurrences in order for ECCMA address elements to fully populate templates.

A second recommendation of this paper is that the CEN element list be opened up to encompass lower level elements from other parts of the world, in such a way as to maintain the integrity of its conceptual foundation. An example of this would be support of U.S. fields such as pre-directionals and post-directionals, which need to be separately identified and separately validated to allow proper handling of a large number of U.S. addresses. For the UPU purposes, it would be possible to do this by turning the "thoroughfare qualifier" element into a construct, with several elements defined within that construct. An alternate approach is to keep "thoroughfare qualifier" as a terminal node element, and allow for synonyms, multiple sequentially defined occurrences, and multiple occurrences of this element to be registered with distinct XML tags and EDI codes. There are certainly other examples of this sort of procedure that apply to other countries. It may be a necessary step in turning the CEN element list into a definitive UPU element list with worldwide scope.

The registry approach above would have further advantages in allowing interoperability of multiple element lists based on mapping the lowest levels to the UPU element list, which would function as a kind of address element ontology. In addition to a registry for tags and codes, there could also be a registry for rendition instructions, which could be a primary worldwide resource for the collection and dissemination of this information, which is critical to the final presentation of the addresses on mail pieces.

A third recommendation of this paper is to move forward on the Status 0 application, with the scope of the proposed standard including elements, templates, and rendition instructions, and to carry out limited testing in several countries from diverse parts of the world once the application has been approved. The general idea is to use the CEN conceptual structure, adding additional elements for which a sufficient justification can be shown, making sure that all of these are supported by ECCMA codes, defining the templates in PATDL or another formal or natural language, documenting the rendition instructions needed for final presentation, and carrying out the test plan as presented separately. Once the tests have been successfully completed, the components can be reviewed once again to determine their adequacy for large scale and more comprehensive testing, and changes can be made along the way to the higher levels in the UPU standardization process.

Joe Lubenow
Industry Co-Chair
UPU DMAB Address Management Project Team
July 8, 2002