Ann Reg Sci (1999) 33:197–212

The Annals of
Regional Science

Location expression standards for ITS: Testing the LRMS Cross Street Profile

Val Noronha[1], Michael Goodchild[1], Richard Church[1], Pete Fohl[2]

[1] Vehicle Intelligence and Transportation Analysis Laboratory, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA 93106-4060, USA
(e-mail: noronha, good, church@ncgia.ucsb.edu)
[2] ISERA Group, 5370 Hollister Ave #2, Santa Barbara, CA 93111, USA
(e-mail: pete@isera.com)

Abstract. Location can be expressed in a number of ways: coordinates, street addresses, landmark references, linear references, grid references, etc. The Cross Streets Profile is part of a location messaging specification developed for use in Intelligent Transportation Systems (ITS). It describes a location essentially in terms of street names. This paper presents an evaluation of the Cross Streets Profile based upon lab and field tests using commercial street network databases in the County of Santa Barbara. Results vary widely, depending upon database quality which in turn is dependent on geography (major streets vs minor streets, rural vs urban). Success rates can be improved with intelligent algorithms. Future directions to improve the quality of location messaging are discussed.

Introduction

A fundamental prerequisite to free exchange of geographic data is the ability to express a location, determined with respect to one map database, meaningfully with respect to another database. Location is always measured and expressed with respect to some frame of spatial or topological reference. A street address is explicitly relative, relying on named objects for orientation; even "absolute" latitude-longitude coordinates are expressed relative to a geodetic framework. Within a single database, locations can be expressed relative to each other with great reliability. Consider a row of houses, for which coordinates have been captured using GPS. When overlaid on the street centerline, the houses may plot on different sides of the street due to inaccuracies in the house or centerline coordinates. Alternately the houses could be topologically linked to the centerline, as a set of left-side addresses and offsets. This relativistic representation skirts the coordinate inaccuracy problem, and is independent of map scale.

For a location to be portable to another database, location must be speci-

All correspondence to Dr. Val Noronha, Department of Geography, UCSB, Santa Barbara, CA 93106-4060, USA

fied using references that are valid across both databases. Traditionally and conceptually, GIS has relied upon the coordinate framework as the common reference, but this breaks down at larger scales—the relationship between map scale and positional accuracy is explicitly recognized in current map accuracy standardization efforts (e.g. FGDC 1998). Emerging applications are being conceived at ever increasing map scales, and data accuracy requirements are more stringent than the needs that current databases were designed to satisfy. Positional differences of up to 50 m are routinely encountered between current commercial databases of the same area; errors of 100–200 m are not uncommon in rural areas or on winding roads (VITAL 1997a). As a result, the latitude and longitude of say a motel, captured with respect to database A, could plot incorrectly relative to database B—the error may range from the wrong block to the wrong neighborhood or the wrong side of a freeway. Whether or not such a location exchange is acceptable to the user depends on the error tolerance of the application. If coordinate referencing is inadequate, then alternate methods must be explored for large-scale applications.

Locations specified with respect to map objects are similarly susceptible to error. There are errors of inclusion and exclusion in maps. Streets are shown to intersect when in reality they do not. Names are often incorrectly recorded, or coded in different ways. Street names and address ranges may be missing. To communicate any location using objects such as cross streets requires that each database contain such information accurately.

Intelligent Transportation Systems (ITS) is an important emerging application area for geographic data exchange. Whether in highway incident management, Advanced Traveler Information Systems (ATIS) or Traffic Management, ITS requires the continuous communication of data about event locations and traffic flows. Trucks transmit their location to fleet headquarters; Traffic Management Centers (TMCs) monitor freeway conditions and issue congestion reports; intelligent vehicles involved in accidents call emergency services on the driver's behalf, if the driver has been incapacitated.

Location expression and exchange (LX) remains one of the major outstanding areas in ITS where accuracy is important, and adequate standards have yet to emerge. Consider a serious accident on a complex freeway interchange. The incident location is captured by a Global Positioning System (GPS) in the vehicle, and relayed to the 911 (emergency response) center. With respect to the 911 database it appears on the wrong ramp, resulting in a 5–10 minute delay in response. Or consider a message regarding traffic congestion, broadcast by FM subcarrier to suitably equipped vehicles. Relative to some databases, the message references a service road rather than the freeway, causing in-vehicle navigation systems unwittingly to recommend the congested freeway route. Three-dimensional coordinates, which are usually available from GPS, may help in some cases, such as distinguishing between ramps in an interchange. Currently few street databases are populated with the third dimension, and there is little reason to believe that elevation values would not also be subject to precision and accuracy limitations, at least in the short term. These location referencing problems are a serious impediment to the implementation of ITS technology.

There are different ways to deal with map inaccuracy, principally: (a) to establish quality and fitness-for-use standards; (b) to create a publicly available framework upon which vendors may build added value. Strategy (a) could be perceived by the private sector as heavy-handed, unless it were accompanied by appropriate incentives and public investments. An example of (b) is the TIGER national database in the United States. Such a framework is expensive to build, and more so to maintain as user requirements evolve.

One strategy proposed for ITS is to establish a broad standard of LX methods—coordinates, street addresses, cross street names, landmarks, linear and grid references. One or more of these methods may be employed for a given situation, depending upon the needs of the communicating parties. The Society of Automotive Engineers (SAE) in the United States is considering the Location Reference Messaging Specification (LRMS) (Goodwin et al. 1996) as the basis of a standard for LX needs in ITS (SAE 1998). In LRMS parlance, each approach to location expression—coordinates, linear references, etc—is formalized into a communications specification at the bit level and termed a "profile." In accommodating a variety of user groups via application-specific profiles, the LRMS takes a reactive rather than a prescriptive approach, one that is palatable both to user groups and to map database vendors.

This paper reports on the testing of the Cross Streets Profile (XSP), perhaps the most prominent in the LRMS suite. Although the profile test is somewhat narrow in its focus, this paper also illustrates the general kinds of problems encountered in databases, remedial approaches, and their likelihood of success. More generally, this research is a step towards addressing the universal interoperability problem that exists between data gathered by different methods, for different purposes and markets. Research into the aspects of the problem, and their solutions, is valuable to anybody facing multiple data sources, the problems of data sharing between them, and the design of appropriate standards. The Cross Streets Profile is described first; the following sections outline the test approach and results. The Conclusions discuss standardization issues, alternative methods of location expression and error reduction.

The Cross Streets Profile

A location message is a stream of digital data that describes a location. For example, a 911 center summons an ambulance to an incident by means of a digital message transmitted to a mobile display terminal or computer in the vehicle; or a crashed vehicle with an unconscious driver automatically senses a problem and reports its position to emergency services. The message could involve human-machine interaction, or it could be entirely machine-to-machine. Because communication takes place over limited wireless bandwidth, the message must be compact, yet maximally informative.

The XSP (Fig. 1), as the term implies, "transfers" a street segment using names of intersecting streets. Specifically, the XSP references a location in terms of (a) the "on-street" on which it occurs, (b) the from-cross-street and to-cross-street defining the segment, and (c) the distance offset along the segment, measured from the crossing with the from-street, and expressed as either an absolute distance or a proportion of the segment length. Crossings do not have to be navigable intersections: an overpass is an acceptable cross street—this is important because current GIS software may blur the distinction between overpasses and physical intersections. The XSP allows for a second offset, to transfer a section of road rather than just a point.

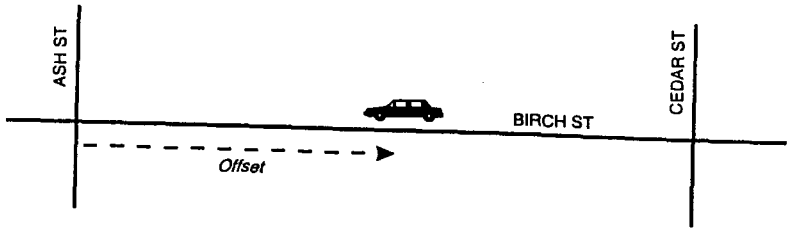

Fig. 1. The Cross Streets Profile expresses the vehicle location as an offset distance along Birch between Ash and Cedar

To the extent that XSP testing consists in part of matching street names against a reference database, there is some commonality with the problem of lexical matching in general (e.g. the Soundex technique, developed for analog computers and patented by Odell and Russell in 1918; Knuth (1973) discussed methods for digital computers) and the associated process of geocoding (deriving coordinates from a street address—a technique used widely in market research and other applications). It is well known that geocoding match rates can be low. Kim and Nitz (1994) report success rates of 35% matching names and addresses in vehicle registration and driver's license lists. Algorithms have been developed by researchers at Statistics Canada (Fellegi and Sunter 1969) and the U.S. Bureau of the Census (Jaro 1989), to deal with these problems, and user-friendly software solutions for geocoding are now commercially available.

There is much more to the XSP test than simple geocoding: (a) the XSP employs three street names, not one; (b) the cross streets are spatially and topologically related, introducing special problems and approaches to solution; (c) sampling must consider all potential vehicle locations: ITS is particularly interested in freeways and ramps, which do not occur in address databases; (d) derivation of coordinates from street names is only a small part of the problem; the vehicle begins with its coordinates from GPS, translates them into street names by snapping to the nearest street and inferring the cross-streets, the names are referenced against a different database at the receiving end, and coordinates are then recovered for cartographic representation.

Test issues and experiment design

The XSP may be applied in a variety of circumstances, from emergency management to relatively non-urgent ATIS enquiries (e.g. a traveler looking for a motel). Success rates with regard to any given implementation are not universally meaningful. Our approach therefore is to isolate the various components of the profile and to test them individually. Testing is thus most useful in that it points out areas in which to focus further development effort.

There are two aspects to testing a profile: one is to examine its inherent merit, to identify pathological cases where the profile could lead to ambiguity or failure; the other is to subject the profile to practical tests, to estimate the likely success of messaging in the context of currently available databases and typical application scenarios.

Preliminary critique

As opposed to coordinates, which constitute an absolute means of location expression, XSP is inherently topological, relying on connective relationships between streets. There are two strong arguments in its favor: (a) it is intuitive, being analogous to conversational driving directions; (b) it is insensitive to minor positional errors in databases. On the negative side, it is verbose, requiring 3 alphabetic strings—say 25 characters or 200 bits in most instances—to store each location.

An obvious potential for error exists in the XSP specification. Since there is no geographic reference other than a triad of street names, a matching triad may theoretically be found anywhere in the world. During the early stages of our test effort, the XSP was being proposed in international circles as *the* basis of a general purpose location referencing protocol for ITS communications. In that context we recommended that the profile be expanded to include a pair of coordinates. This change was later incorporated in the profile as an optional expression (leading to the specification in Table 1), and a second round of testing followed.

Inclusion of coordinates considerably strengthens the XSP. At the cost of about 70 additional bits per point, it employs two entirely independent and therefore redundant LX approaches: referential and absolute. If the receiving party finds the two expressions to be in conflict, an error is known to be present. The enhanced XSP therefore incorporates a measure of messaging

Table 1. The Cross Streets Profile (from SAE Information Report J2374, May 1998)

Bit	Content	Values/Range
0–3	Start Code	0100
4–7	Pad	
8–15	On Street Byte Count	0–255
16–23	From Street Byte Count	0–255
24–31	To Street Byte Count	0–255
32-variable	On Street Name	(On Street Byte Count) ASCII characters of name
variable	From Street Name	(From Street Byte Count) ASCII characters of name
variable	To Street Name	(To Street Byte Count) ASCII characters of name
3 bit	Horizontal Datum	000 = WGS-84; 001 = WGS-84+EGM-96; 010 = NAD83; 011 = NAD27; 100-111 = Other
32 bit	Start Longitude	−180,000,000 → +180,000,000 microdegrees
32 bit	Start Latitude	−90,000,000 → +90,000,000 microdegrees
32 bit	End Longitude	−180,000,000 → +180,000,000 microdegrees
32 bit	End Latitude	−90,000,000 → +90,000,000 microdegrees
3 bit	Vertical Datum	000 = WGS-84; 001 = NAVD-88; 010 = Vertical Level Code; 011-111 = Other
16 bit	Start Altitude	−8191 → +57344 decimeter OR −7 → +7 Level Code
16 bit	End Altitude	−8191 → +57344 decimeter OR −7 → +7 Level Code
24 bit	Offset1	−8,388,607 → +8,388,607 decimeters
24 bit	Offset2	−8,388,607 → +8,388,607 decimeters
2 bit	Side	00 = right-hand; 01 = left-hand

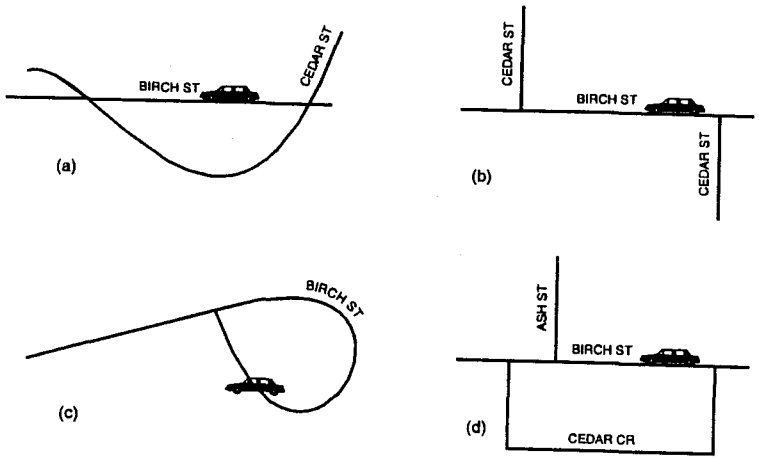

Fig. 2. a–d. XSP expression of location is ambiguous when these street configurations are encountered **(a)** multiple crossings; **(b)** significant jog in cross street; **(c)** self-crossing loop; **(d)** crescent.

confidence. Moreover, if the street names cannot be processed for some reason (e.g. cannot be found in the target database), or the coordinates point to a location in the sea, the other XSP component still works, albeit with reduced confidence.

To be a general messaging standard, the XSP should be able to accommodate any arrangement of streets and naming conventions. Preliminary examination shows that a number of common street configurations and municipal practices lead to ambiguities (Fig. 2). Coordinates can theoretically assist in resolving these if their accuracy, and the accuracy of the reference databases, are sufficient.

Implementation and outcomes

In an ITS implementation of the XSP, a location is conveyed from a sending party (say a TMC) to a receiving party (the navigation system in a vehicle traveling down a freeway). In this example the communication is typically one-way, by FM subcarrier broadcast or similar technology. An alternative scenario is where the vehicle is equipped for transmission (e.g. cellular modem), enabling two-way communication. After the message has been transmitted, there are three possible outcomes at the receiving end:

- Success—the message is successfully received and interpreted; or
- Failure—the message cannot be interpreted, perhaps because the street is missing in the target database, or the coordinate points to an improbable location such as the sea; or
- Error—the message is interpreted, but unknown to the receiver, it points to the wrong location.

Table 2. Two broadly different approaches to interpreting an XSP message

Implementation	Possible outcomes
A Given the coordinate, search its neighborhood (say a 500 m window), or an expanding window, for a matching triad of names (on-street, cross-street 1, cross-street 2).	Degrees of name agreement: 1. Perfect match found 2. Probable match found on fuzzy search 3. Unreliable match found 4. No match found
B Given the cross street names, resolve the location. Independently, snap coordinate to target database. Check two locations for agreement.	Degrees of name agreement (as above) Degrees of distance agreement Degrees of corroboration between name-based and coordinate-based solutions.

In a two-way scenario failure is preferable to error, because the receiver may request a re-send. In a one-way scenario one may take the position that some information is better than no information, and some error may be tolerable. Neither the XSP nor any of the associated SAE communications standards currently in circulation provide for measures to be taken in the event of message failure.

The XSP specification does not prescribe how a message is to be interpreted by the receiver. In the event of a conflict between coordinates and street names, a method for resolution is not prescribed; that is up to users. There are at least two approaches to interpretation (Table 2). One is to constrain the search of the target database to the neighborhood of the given coordinate. The other is to conduct two independent searches of the target database, one for matching coordinates, the other for matching names; and to test for agreement of the results. Approach #1 is more efficient, but assumes that the coordinate is correct or dominant; depending on implementation details, it could require that the cross street intersections lie within the neighborhood of the transferred point location. Approach #2 is more flexible and robust; it allows for the discovery of a name match whether or not it is within the vicinity of the coordinate.

The testing reported in this paper follows approach #2, not necessarily because it is thought to be superior, but because the dual independent search for name and coordinate match illuminates a wider variety of database problems and outcomes.

Metrics

When the XSP is employed without coordinates, there are two core evaluation issues: first, how well does the triad of cross street names identify the correct street segment; and second, assuming the segment is correctly identified, how accurately does the distance offset represent the location of the point on the segment. The first component is critical; no matter how well the offset works, reference to the wrong street usually constitutes an intolerable error. On the other hand, there is varying tolerance to offset error, depending on the application. There is a third, composite issue—2-dimensional error—where again there is varying tolerance. For example, if emergency personnel are summoned to a large fire, the flames are so visible that an error of even two city

blocks is of little consequence in *locating* the incident, whereas the same error could pose a serious problem for those routing a relief vehicle to it.

Offset error is central to another LRMS profile, the Linear Referencing Profile, also under study at VITAL. A comprehensive treatment of it is not possible here due to space limitations. This paper focuses on segment identification, with some attention to two-dimensional error.

The accuracy of segment identification can be measured by a "hit rate" or the empirical likelihood of success. A hit is nominally defined as any successful interpretation of the message with reference to the target database. Spurious hits can be recorded if multiple instances of the name triad exist in the target database.

With the inclusion of coordinates, new issues arise. The test of success in the names-only profile is merely the ability to find a matching triad. Only a controlled test can determine whether or not that is the intended instance of the triad. With the enhanced XSP the test of success is whether coordinates corroborate or challenge the cross street result. This test need not be restricted to the lab; it can and should be performed in normal use. Our study therefore focuses on aspects and statistics related to message interpretation, that would not normally be possible in operational use. It is a commentary on the process, as much as a test of the system.

As outlined earlier, there are two broad approaches to the use of coordinates: (1) an initial seed point for a search over a widening radius, and (2) a separate search method, independent of street names. At the implementation level, coordinates can further be used as (3) a tie-breaker between two matching triads, or (4) a fallback if location cannot be determined from names alone. Our implementation takes approach #2 with a distance criterion to judge success; it counts the instances in which operations (3) and (4) are necessary.

A separate series of tests examines the converse process, using coordinates alone to make the match, verifying the accuracy of the transfer by name matching. The algorithm does not search outward for a better match as in approach (1) above; it tests only the nearest street segment.

Data sets

Testing was based on five major commercial street databases covering the County of Santa Barbara. The County was selected as the study area because it covers a variety of terrain and consequently has a broad range of road densities and styles, from freeways to winding mountain and ranch roads. It would not be appropriate to identify data vendors, first, because the test results may be interpreted as reflecting their strengths or weaknesses (some databases were not developed to support ITS applications); and secondly because we view interoperability as a problem in its own right, that exists despite the best efforts of vendors.

Data were acquired at about the same time (May 1997), directly from vendors. Streets were extracted, and all coordinates transformed to a common data structure, datum and projection (NAD83, UTM). For the second round of tests (coordinates included) pseudo-nodes were eliminated, and dead-ends specially coded to distinguish them from intersections with unnamed cross streets.

The density of the street network varies considerably between databases. Some include private ranch roads or driveways within residential or commercial complexes, others do not. There are also errors in classification, with bike paths and railway tracks being coded as streets. One database was still under development by the vendor at the time of testing; outside of the City of Santa Barbara it contained only a bare skeleton of major highways. This database could be used only for some tests. The databases used for the majority of tests are coded A, B, E and F.

Data sets also vary in terms of attributes. Some databases are designed specifically for navigation, and contain precise data on distance, directional restrictions, road classification, speed limits and travel times. Other databases are designed for traffic control, and do not require these attributes; yet others are primarily used to geocode customer addresses for marketing studies. It could be argued that it is inappropriate to study ITS interoperability between such disparate databases. While database suitability is well understood by those intimate with the corporate evolution of specific products, users often choose data products inappropriate to their needs. This is why interoperability is considered a major problem in the ITS world.

The attribute of greatest interest to the XSP is the street name. Curiously, no two databases structure this in exactly the same way. In the case of "Birch Street," it is easy to parse the expression into a proper name ("Birch") and type ("Street"). Directional prefixes and suffixes are recognized by some vendors; only one vendor accommodates a street type prefix (as in "Avenida Redondo"), common in French and Spanish areas. There is disagreement about highways, e.g. "Hwy 101" vs "Rte 101." Even street type abbreviations, for which the U.S. Postal Service has prescribed standards, are not universal. Another serious problem is highway aliases (e.g. US-101 is variously known as Ventura Freeway and El Camino Real). Not all databases have alias fields, some use numbers rather than names, others follow the opposite practice.

Naturally there are human errors, such as "Hollister Avenue" coded as "Hollister Street." Underlying data structures and database design interfaces can be designed to enforce consistency and to minimize human error, but this level of quality control was found to be absent in some databases. A long street might have different versions of its name applied to various segments due to typographic errors, making it difficult to link the components into a continuous path.

Fortunately, as noted earlier, there are solutions to these problems. The general problem of fuzzy lexical matching—and street name matching in particular—is now well addressed in commercial software, using a variety of techniques from "blocking" and transposition to phonetics and probabilities (Jaro 1989). These products can equate a search key of, for example, "Holiister Street" (mis-spelt, with street type error) with a database entry of "Hollister Ave." Note that these products are typically used to match a poorly specified key (e.g. an unverified address provided by a telephone subscriber) against a relatively clean database, not to match one relatively clean database against another—the problems are slightly different. Due to time constraints we were not able to make use of such software, but developed a rough substitute instead.

Another troublesome aspect of data sets is their effective scale and resolution. A divided city road is stored by one vendor as dual centerlines; other vendors represent it as a single line. Similarly traffic circles may be generalized

into regular 4-way street intersections, and channelized "slip" turn lanes may or may not be shown. These are not strictly speaking errors in databases; they are the inevitable consequence of varying vendor philosophies and levels of generalization. But a LX method that relies on topological relationships clearly falters due to these discrepancies.

Sampling

Two samples are tested. The first is a lab-generated set of points chosen at regular distance intervals along the street network for each database. Tests are based on 1000 points; sampling density per kilometre may differ slightly because of differences in street length and network density. All points in the 1-dimensional network space of the county are equally likely to be selected; this represents one sampling extreme, appropriate to worst-case service provision such as EMS, where an incident may occur at any point. The second sample is a set of 54 points surveyed by VITAL's research vehicle using differential GPS. Points were selected at recoverable physical locations (e.g. motel entrances) along major publicly accessible streets in the immediate vicinity of the City of Santa Barbara. The locations were documented by notes and photographs. This sample represents another extreme, unrepresentative of rural needs, but more appropriate to ITS.

For lab and field tests, each test point is expressed as a XSP message in a source database, and transferred to each other database in turn (the "target" database). Success rates or other measures are computed from each database to every other database. For a study of 4 databases, this amounts to 12 pairwise combinations (i.e. source database, target database). The mean and other statistics, when quoted below, are summaries of these 12 observations.

Test results

A serious problem with the data is the large proportion of records with blank street name fields. In the study area, 20 to 45% of all records, accounting for 40 to 60% of road length, have blank names (Table 3). Since a XSP triad requires three populated name fields, many transfer attempts are rejected on this point alone. The column "% blank triads" shows that in up to 31% of all possible triads, all three names are blank. On examination, many of the blanks are private ranch roads which, it could be argued, are of limited significance for ITS. But many others are highway ramps, which are important for ITS and emergency reporting. On average, valid triads could be formed in

Table 3. Incidence of blank street names

Database	Total records	Blank records	% Blank records	% Blank triads	Formation of valid triads (%) Lab sample	Field sample
A	35483	15518	44	31	11	39
B	32687	10637	33	19	18	63
E	20067	3844	19	10	29	67
F	30000	12339	41	27	15	52

just 18% of all attempts with lab-generated points, and 55% with field sampled points (i.e. points on major public streets).

An initial series of tests matched name triads from a source database back to the source database. This was designed to be a "shakedown" test of the computer code, which should have produced a 100% hit rate. The rate actually achieved was 90–94% of valid triads; remaining transfers failed due to topological ambiguities (Fig. 2) and multiple hits.

Predictably against the backdrop of database discrepancies, the results of matching across databases were not impressive. Using the original XSP specification (no coordinates), the success rates for non-blank transfers were 19% (lab) and 25% (field). The low rates were due to lexical mismatches and topological problems outlined earlier. When failures due to blanks were included, the net success rate was just 3% (lab) and 14% (field).

These numbers were obtained using basic algorithms that searched for exact lexical matches and allowed first-order intersections only. Clearly the XSP would fail when implemented in this manner, but might be more successful if supported by intelligent software at the transmitting and receiving ends. A number of improvements were made for the second round of testing:

- Databases were pre-processed to eliminate pseudo-nodes (bivalent nodes); dead-end streets (univalent nodes) were specially coded.
- Fuzzy lexical matching was implemented, and equivalence tables were developed to accommodate some of the vendor differences in highway labeling (e.g. to equate "US Hwy 101" with "US-101").
- Judgement criteria were relaxed to admit blank streets under some circumstances; apparent successes due to blanks were counted separately.
- Some tests were run on streets classified as major, rather than on the small selection of field-sampled points.
- To cope with the blank streets problem, an intelligent algorithm was developed to search outwards from the transfer point, for 2nd, 3rd or nth-order cross streets that might be major or non-blank.

The following sections report on these tests.

Coordinates alone

This test series composed XSP messages using coordinates alone. The test point was snapped to the nearest segment in the target database, subject to a maximum distance of 300 m. The test of messaging success is whether or not the point snaps to the intended street, as judged by comparing street names. However, the discussion above shows that there are numerous problems with matching street names. A series of rules was therefore devised, whereby matches were characterized as "likely," "possible" or "unlikely," with some blank names accommodated. Matching outcomes were scored as follows:

- Hit: Score 2 if the source and target street names match
- Miss: Score 0 if there is a mismatch between two non-blank names, *or* if the source is blank but the target *can* be found elsewhere in the source file, *or* the target is blank but the source *can* be found elsewhere in the target file.
- Unsure: Score 1 otherwise.

Table 4. Results of Coordinates-alone tests

	Lab sample			Field sample		
	Mean	Min	Max	Mean	Min	Max
Likely	11%	2%	18%	24%	6%	46%
Possible	61%	52%	70%	26%	22%	31%
Unlikely	28%	18%	45%	50%	30%	65%

Based on these scores, outcomes of the transfer were categorized as likely, possible or unlikely, using a list of rules. The test results appear in Table 4.

This test is not a fair representation of practical application of the XSP. It merely measures the effectiveness of coordinates used alone, and is biased by name problems in the databases.

Name first

The name-first approach processes cross street names, and uses coordinates in a secondary capacity to break ties in the event of multiple hits, or to check the accuracy of a match. Once a hit is achieved, coordinates are used to measure Euclidean distance to the transferred point, as a measure of accuracy. This approach ensures that all matching triads are examined, no matter how far they are from the known coordinate.

The algorithm for street name processing is described in Fig. 3. Nine outcomes were identified, and for each test the number of instances of each outcome recorded. The outcome counts represent the likely success at different levels of sophistication. For example, with no fuzzy name matching, Bin 1 rates can be expected. Success of the algorithm is measured by (a) the proportion of unambiguous transfers (Bins 1,2,5,6), stalemates (Bins 3,4,7,8) and fallbacks (Bin 9), and (b) corroboration by coordinates, i.e. distance from the source coordinate to the transfer point on the target street—manual inspection indicated that discrepancies greater than 30 m generally indicated an incorrect transfer. These results are quoted in Table 5. One of the test databases is known in the industry to have been derived as a value-added product from another. Positional correspondence is therefore extremely good between these two databases, causing the minimum distance results in Table 5 to be close to zero.

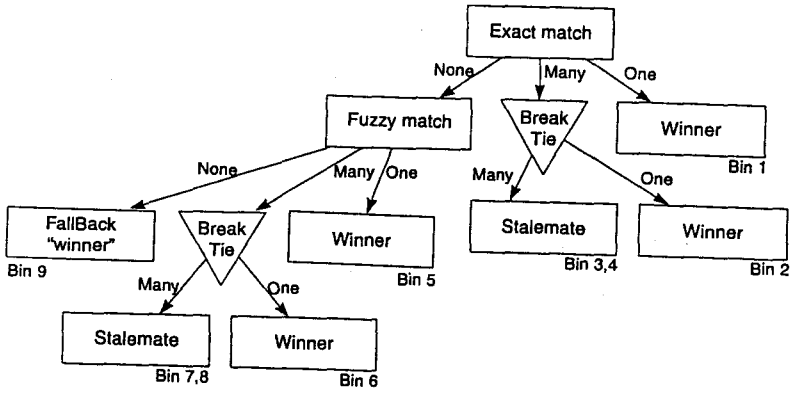

Fig. 3. Flowchart for name processing

Table 5. Results of name-first tests using basic algorithm. Note that Min and Max columns do not add up to 100% because they may reference different pairs of databases

County of SBA	All streets			Major sreets		
	Mean	Min	Max	Mean	Min	Max
Unambiguous successes (1,2,5,6)	43%	19%	74%	22%	2%	80%
Stalemates (Bins 3,4,7,8)	21%	7%	39%	9%	0%	27%
Fallbacks (Bin 9)	35%	16%	45%	68%	38%	91%
Distance discrepancy < 30 m	66%	37%	96%	82%	52%	97%

Major streets were stratified for separate treatment in this test, rather than using the limited sample of field points. Street classification systems vary significantly between vendors—one vendor uses 6 categories, another uses about 40. Sample points were generated along streets characterized as major in the source database, and transferred to *any* matching street (not necessarily major) in the target database. This closely mimics a real world application scenario: incidents tend to occur on major streets, but receiving databases would typically have all streets on file.

On average, results were not as good with major streets, for two reasons. First, while incidents are restricted to major streets, cross streets may still be minor unnamed roads. Second, highway aliases (e.g. "Hwy 101" vs "Ventura Freeway") remain a problem. This can be overcome with lookup tables; we did not develop these for testing, but commercial information service providers might do so and achieve better results. The success rate between the most compatible pair of databases—80%—was higher for major streets than for all streets.

Intelligent matching

A final set of tests was run using a Nearest Qualifying Cross Street (NQX) algorithm. The algorithm builds a XSP triad by searching out from the

Table 6. Results of name-first tests using NQX algorithm on entire county of Santa Barbara

County of SBA	All streets			Major streets		
	Mean	Min	Max	Mean	Min	Max
Unambiguous successes (1,2,5,6)	45%	20%	74%	18%	0%	61%
Stalemates (Bins 3,4,7,8)	22%	7%	40%	11%	0%	34%
Fallbacks (Bin 9)	33%	15%	44%	72%	45%	98%
Distance discrepancy < 30 m	65%	36%	96%	81%	47%	100%

Table 7. NQX algorithm applied to urban Santa Barbara

City of SBA	Major streets		
	Mean	Min	Max
Unambiguous successes (1,2,5,6)	23%	1%	71%
Stalemates (Bins 3,4,7,8)	26%	0%	84%
Fallbacks (Bin 9)	51%	23%	93%
Distance discrepancy < 30 m	84%	66%	100%

transfer point for qualifying names (non-blank; major if applicable) in the source database, creating a chain rather than a single link. At the receiving end the algorithms searches for the two intersections {On, From} and {On, To}—blank names are not admitted at all. It examines all possible paths between these two intersections, constraining the search to links carrying the On-street name (in theory there should be only one path, but due to database errors and odd municipal practices, such as forking streets with the same name, multiple paths are sometimes encountered). It selects the longest such path to be the target segment. In effect, this approach ignores blank streets; when applied to the major streets LX problem it forgives some topological discrepancies (minor intervening cross-streets in one database and not the other). NQX was implemented in conjunction with other matching processes, in the sequence (1) exact match, (2) NQX exact match, (3) fuzzy match, (4) NQX fuzzy match.

NQX produced just a 1–2% improvement over previous results, and some results appear to have worsened. Time constraints on the research did not permit further development; however the principles of NQX appear sound, and better results can undoubtedly be achieved with additional effort.

Urban areas would be expected to produce superior results because one would expect better database quality. In our tests of the urbanized areas of Santa Barbara and Goleta, this was not the case (Table 7). However, independent XSP tests were carried out by Haas et al. (1998), using San Francisco and Santa Barbara data. Experimental procedures were slightly different, but results for Santa Barbara were substantially similar, while better match rates were recorded for San Francisco. There are two explanations for this. One is that the San Francisco study area is uniformly urban, more so than Santa Barbara; a second explanation is that the San Francisco Bay area is the center of operations for a number of data vendors, perhaps resulting in a tradition of better data quality.

For the sake of brevity this paper quotes only the principal results of the XSP tests. Complete tables of results are quoted in the formal test reports (VITAL 1997b, 1998).

Conclusions

The XSP evaluation cannot be summarized in a single figure or even a descriptive sentence. Success is a function of numerous factors, from inherent profile effectiveness to municipal practices, geography and database accuracy. Since the XSP does not specify how a message is composed or processed at the receiving end, implementation is a controlling factor in its success. With so many dimensions to the tests and results, each potential user group will need to focus on aspects appropriate to its needs.

Even with intelligent algorithms, transfer success rates are not as high as the ITS industry, and emergency management services (EMS) in particular, require. Low success rates are largely due to inconsistent database quality, particularly the high incidence of blanks, absence of aliases in many databases, and non-standard name parsing and abbreviations.

The following are constructive recommendations for national-level activities that would lead to better results.

1. The ideal long-term course of action is to re-survey the national street network to uniform quality standards. Piecemeal efforts are already underway. Several municipalities have integrated GIS programs in operation, with varying degrees of coordination between federal, state, county and private agencies. Outstanding hurdles are (a) the technical difficulty of finding a common quality standard that suits the needs of all stakeholders at a justifiable cost, and (b) management challenges to coordinate this activity at a national scale. Even if re-survey is publicly funded, commercial vendors will need to make substantial investments in data reorganization and conflation of nationwide databases. There are two shorter-term measures: standardization of databases, and the ITS Datum.
2. Standardization of databases: Messaging could be simplified if vendors would establish and populate alias name fields, and comply with basic standards in street naming, in particular, highway and ramp nomenclature, field separation and abbreviation. Standardization of other aspects such as classification and coverage, are desirable, but may not be readily achievable in the short term.
3. The ITS Datum (Siegel et al. 1996) is a mid- to long-term strategy that could potentially alleviate many current messaging problems, provide an evolutionary framework for a high-quality national database, and offer a mechanism for continuing update of highway-related coordinates and attributes, that would survive future construction and changes in geodetic datums. The ITS Datum would consist of a set of precisely surveyed points at significant locations such as street intersections. Disagreements between databases would be documented and/or modeled, and corrections may be applied statically (i.e. a vendor elects to make the required changes at source) or in real time. This approach would not necessarily require a complete re-survey of the street network. The Datum would be established only to the density and accuracy required to enable map corrections to satisfy the tolerances of applications. Conceptual design and preliminary testing of the ITS Datum are underway at VITAL (Funk et al. 1998; Church et al. 1998); improvement in XSP success may be one of several measures of its ultimate effectiveness.

Finally there are problems with the XSP specification itself. First, it lacks metadata content. For mission critical applications, a message should express the level of certainty with which the location is known, and the receiver should interpret the message in the context of that quality statement. Second, the XSP assumes that communication is one-way, and that the message is satisfactorily received. With two-way wireless TCP/IP communications technology now available (e.g. cellular digital packet data or CDPD, capable of 19,200

bps), the receiver can examine a message for consistency (e.g. cross streets and coordinates must agree within a certain tolerance) and request re-transmission using other profiles if necessary. These areas of development are currently under investigation at VITAL.

Acknowledgements. This research was supported by the United States Department of Transportation, Federal Highway Administration, Contract DTFH61-91-Y-30066, and the California Department of Transportation, Testbed Center for Interoperability, Interagency Agreement 65V250. The Vehicle Intelligence and Transportation Analysis Laboratory (VITAL) is a testbed for ITS spatial data interoperability. VITAL performed this research under contract to Viggen Corporation, Interop Division.

References

Church R, Curtin K, Fohl P, Funk C, Goodchild M, Kyriakidis P, Noronha V (1998) Positional distortion in geographic data sets as a barrier to interoperation. Proceedings, ACSM Baltimore [document available at www.ncgia.ucsb.edu/vital]

Fellegi IP, Sunter A (1969) A Theory for Record Linkage. Journal of the American Statistical Association 64 (328):1183–1210

FGDC (1998) Content Standard for Digital Geospatial Metadata. United States Geological Survey, Reston VA. Federal Geographic Data Committee, Metadata Ad Hoc Working Group. FGDC-STD-001-1998.

Funk C, Curtin K, Goodchild M, Montello D, Noronha V (1998) Formulation and test of a model of positional distortion fields. Third International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Quebec City [URL www.ncgia.ucsb.edu/vital]

Goodwin CWH, Siegel D, Gordon S (1996) Location Reference Message Specification: Final Design. Task B: Spatial Data Interoperability Protocol For ITS Project, United States Department of Transportation, Federal Highway Administration Office of Safety and Traffic Operations, ITS Research Division, Contract 61-94-Y-00001, Draft, June 28, 1996

Haas RP, Lau J, Goodwin C, Gordon S (1998) Location Referencing Message Specification Report. Version 1.0, July 23, 1998

Jaro M (1989) Advances in Records Linkage Methodology as Applied to Matching the 1985 Census of Tampa. Journal of the American Statistical Association 84(406):414–420

Kim K, Nitz L (1994) Application of Automated Records Linkage Software in Traffic Records Analysis. Transportation Research Record 1467:50–55.

Knuth DE (1973) The Art of Computer Programming, Volume 3: Sorting and Searching. Addison Wesley, Reading, MA

SAE (1998) Surface Vehicle Information Report – Location Referencing Message Specification. Society of Automotive Engineers, Information Report J2374

Siegel D, Goodwin C, Gordon S (1996) ITS Datum Final Design Report. Task C: Spatial Data Interoperability Protocol For ITS Project, United States Department of Transportation, Federal Highway Administration Office of Safety and Traffic Operations, ITS Research Division, Contract 61-94-Y-00001, Review Draft, June 28, 1996

VITAL (1997a) Interoperability of Map Databases – Development of Experimental Infrastructure. California Department of Transportation, Test Center for Interoperability, Interagency Agreement 65V250. Draft Final Report

VITAL (1997b) The Cross Streets Profile – Technical Evaluation. United States Department of Transportation, FHWA Contract DTFH61-91-Y-30066, Draft Phase I Report. [URL www.ncgia.ucsb.edu/vital]

VITAL (1998) The Cross Streets Profile with Coordinates – Technical Evaluation. United States Department of Transportation, FHWA Contract DTFH61-91-Y-30066, Final Report. [URL www.ncgia.ucsb.edu/vital]

Ann Reg Sci (1999) 33:213–232

The Annals of
Regional Science

Semantic interoperability: A central issue for sharing geographic information

Francis Harvey[1]*, Werner Kuhn[2], Hardy Pundt[2], Yaser Bishr[2], Catharina Riedemann[2]

[1] Department of Geography, University of Kentucky, Lexington, KY 40506-0027, USA
(e-mail: fharvey@pop.uky.edu)
[2] Institute for Geoinformatics (IfGI), University of Münster, D-48149 Münster, Germany
(e-mail: kuhn,hardy,bishr,riedema@ifgi.uni-muenster.de)

Abstract. Technical interoperability has provided geographic information communities with substantial improvements for constructing GIS capable of very low friction and dynamic data exchanges. These technical advances stand to provide substantial advantages for sharing geographic information, however reaping these advantages in highly heterogeneous operational and organizational environments requires the understanding and resolution of semantic differences. While the OpenGIS consortium has made important progress on technical interoperability, semantic interoperability still remains an unpassed hurdle for efforts to share geographic information across organizational and institutional boundaries at the local, regional, and other levels. Identifying and resolving semantic interoperability issues is especially pertinent for data sharing and considering future developments of standards. This paper presents an overview of semantic interoperability and through case studies shows the breadth and depth of issues and approaches in different countries and at different levels of organizations. These cases illustrate the importance of developing flexible approaches to practical data sharing problems that merge semantical with technical considerations. Based on our examinations of semantic issues and approaches in ongoing research projects, we propose cognitive, computer science, and socio-technical frameworks for examining semantic interoperability.

1. Semantic interoperability, standards, and data sharing

Interoperability is widely recognized as a new paradigm for joining heterogeneous computer systems into synergistic units that facilitate a more efficient use of geographic information resources. This is part of a more comprehensive enterprise-orientated view of information technology in general. Considerable

Please send all correspondence to F. Harvey
This work is supported by a grant from NCGIA's project Varenius. Portions of this paper were presented at the Interop '99 conference (Zurich, March 1999) and will appear in that proceedings volume.