A comparison of text mining and semantic approaches for integrating national and local habitats data

Semantic accuracy, error or inconstancy?

Alexis Comber
Department of Geography
University of Leicester
Leicester, LE1 7RH, UK
ajc36@le.ac.uk

Andy Lear
Leicester and Rutland
Wildlife Trust
Leicester, LE5 2JJ, UK
alear@lrwt.org.uk

Richard Wadsworth
Centre for Ecology
and Hydrology,
Lancaster, LA1 4AP, UK
rawad@ceh.ac.uk

Abstract—**This paper compares two approaches for integrating data semantics: data primitives and latent semantic analysis. The former is shown to be appropriate when the user has an understanding of the domain. The latter is suitable when the user has no *a priori* knowledge of the domain.**

Keywords: land cover, habitat, integration, semantics

I. INTRODUCTION

There is a long-standing data integration problem in the domain of land cover mapping: how to reconcile semantic differences between classification systems. The origins and reasons for discordant land cover information have been discussed by many workers. The integration of land cover information supports a number of scientific activities related to the identification of landscape changes and fusion of multi-source landscape information. Post classification comparisons of multi-temporal mappings afford the opportunity to identify land cover transitions to support environmental and planning objectives. Additionally there is much interest in being able to use one thematic dataset as the basis for reporting on classes from another classification, for example because of its better spatial coverage.

Overcoming differences in data semantics has been identified as *the* major obstacle to data integration (Harvey et al., 1999; Frank, 2001; Pundt and Bishr, 2002; Comber et al, 2008a) and as Frank (2007a) notes, "In order to achieve interoperability in GIS, the meaning of data must be expressed in a compatible description". A number of different groups of solutions to the integration of classified information (ie post classification integration), that explicitly seek to deal with the data semantics, data meaning and data concepts have been proposed. These include translations to a common reference, and translation to a common semantic framework (eg Di Gregorio and Jansen, 2000; Kavouras and Kokla, 2002; respectively). Data are re-interpreted using the terms of a common framework. These approaches require considerable human (expert) input to specify the class attributes and characteristics that are important and how they relate either to other classes or to the reference system. In one sense these are 'top down' approaches that impose control over the taxonomies, vocabularies and ontologies into which data are recast. However in most cases the 'semantic quality' of any dataset is only relative to the intended application or use (rather than being absolute) and is difficult to anticipate in advance such as are required by top down integration approaches (Comber et al., 2008a). For an example see Alqvist's (2008) critique of the LCCS proposed by Di Gregorio and Jansen (2000). Thus 'bottom up' approaches to integrating data semantics offer the opportunity to accommodate the context of the intended use of the data. That is, interoperability is considered from the user's perspective, who is concerned with what the data categories *mean* and how they relate to other classifications and to other classes in the same categories.

Other work has sought to develop 'bottom up' semantic integration methods including measures of semantic consistency in look up tables. Relationships between classes from different classification systems are encoded. Comber et al., (2004a,b) used expert opinion to describe semantic relationships "expected", "uncertain" and "unexpected" to compare two land cover maps, a problem the data producers warned users was intractable. Wadsworth et al (2005) described land cover attributes in terms of their data primitives to explore inconsistencies between three land cover maps of Siberia and Ahlqvist (2004) the use of conceptual spaces. Data primitives and Conceptual spaces describe the qualities of the data under investigation at the most fundamental level and are scored once for each dataset. They provide information about the underpinning concepts – their meaning -. Land cover data primitives have been addressed by a number of workers. Alqvist (2004) calculated the degree of overlap and conceptual distance between land cover classes using four 'approximation spaces'. Comber (2008) identified a number of conceptual dimensions to separate the concepts of land use and land cover. The data primitive approach identifies the concepts or domains that are important, scores each class within that domain. Scores are then compared across domains for different classes.

A third group of work in this area, seeking to integrate divergent data semantics, has also taken a bottom up approach but seeks to avoid the direct interaction with experts, who may not be available. The papers by Wadsworth et al (2006), Comber et al (2008b) and Wadsworth et al

(2008) show a series of experiments involving different text mining techniques applied to a range of geographical information. The text mining approaches are applied to the knowledge experts have "stored" in written descriptions, as either metadata or as survey memoirs. The text mining approaches have avoided natural language processing because it is a very complex problem, especially when applied to scientific texts. However document categorization and information retrieval methods that make the "bag-of-words" assumption provide a much simpler problem. Wadsworth et al adapted the work of Lin (1997) and Honkela (1997) to look at the similarity between land cover classes rather than documents (Wadsworth et al., 2006). In an attempt to understand the semantic overlaps between the classes they applied Latent Semantic Analysis techniques (PLSA and LDA) (Hofmann 1999a,b; Blei et al 2003). The results identified a number of domains and the semantic concepts that were associated with them.

II. OBJECTIVES

This paper compares two bottom-up approaches to data integration: data primitives using a number of user generated domains and the identification domains using latent semantic analysis techniques. The data primitive approach provides an independent description of the data classes using concepts that are important to the user. It assumes that the user (or at least the person scoring the data in the domains) is familiar with the data. The latent semantic analysis approaches extract the domains and the concepts associated with them. They provide the naïve user with a foothold into the semantic concepts associated with the data. The aim of this work was to explore techniques that will support such naïve users as they are increasingly given access to spatial data through an increasing ubiquity of GIS and eScience infrastructures etc (Comber et al., 2008a). There are then levels of spatial data use and need to understand data semantics that are below the current 'bottom up' approaches to data integration. The context for the work was to support consistent habitat reporting. In the UK regional biodiversity partnerships are responsible for regional habitat reporting to national agencies. However there is considerable within- and between-region variation in all aspects of data collection, data management (e.g. metadata). Local habitat data may be temporally and thematically variable. It is collected by different organisations, often using different habitat classifications, to support varying local priorities and objectives (Gallagher and Calder, 2007; NE and TWT, 2009). The variation in habitat data collection and recording has implications for a number of habitat related activities: it reduces the overall quality and consistency of habitat reporting and is problematic for regional objectives such as identifying habitat opportunities (NE and TWT, 2009).

III. METHODS

The case study demonstrates the two approaches for integrating discordant land cover from satellite imagery with local habitat data collected by field survey. The Leicester and Rutland Wildlife Trust collected the field data and the satellite data was from the Centre for Ecology and Hydrology's forthcoming Land Cover Map 2007. The analysis compared different approaches to generate measures of overlap between 3 local habitat classes (Broadleaved and wet woodland, Mixed Grass, Acid Grass) and 3 national broad habitats (Broadleaved, mixed and yew woodland, Improved Grassland, Neutral Grassland).

A. Data Primitives

Six primitive dimensions were identified as being important to biodiversity and were scored: vegetation structure or canopy cover (Complex to Simple); biomass harvesting (Most to Least); vegetation height (High to Low); soil fertility / agricultural improvement (Most to Least); species richness (Most to Least); size / area (Large to Small).

The dimensions were all scored in a non-ordered qualitative manner using 10 classes, although we note that some could be continuous (e.g. vegetation height, biomass harvesting, area etc). Each class was scored as being present (1) or absent (0) and the measure from Bouchon-Meunier et al (1996) for to non-ordered qualitative domains was applied:

$$O(p_A, p_B) = \frac{\sum \min(p_A, p_B)}{\sum p_B} \quad (1)$$

where $f_{pA}(x)$ and $f_{pB}(x)$ represent the values of concept (classes) A and B at location x in domain p; and pA and pB are the properties of concepts A and B in domain p. The overlap measure can vary from 0 (no overlap) to 1 (class B is a subset of A). Classes will overlap to a different degree in each of the domains. An average score was calculated for the six domains to generate a 'look up table' reporting the overlap between and within different classification schemes. Table 1 shows the example of soil fertility.

TABLE I. SCORING OF SOIL FERTILITY

Type	Soil Fertility	High → Low									
Local	Acid Grass									1	1
Local	Mixed Grass					1	1	1	1	1	
Local	Broadleaved Woodland			1	1	1	1	1			
National	Improved Grassland	1	1	1							
National	Neutral Grassland					1	1	1	1	1	
National	Broadleaved, Mixed and Yew woodland			1	1	1	1	1			

B. Latent Semantic Analysis

The similarity of the classes was examined as in Wadsworth et al (2006) using Probabilistic Latent Semantic Analysis (PLSA). This was proposed by Hofmann (1999a,b) as a "generative" model of latent analysis; the joint probability that a word (w) and document (d) co- occur (P(d,w)) is a function of two conditional probabilities; that the document contains a concept (z) (P(z|d)) and that the word is associated with that concept (P(w|z) (Equation 2)

$$P(d, w) = P(d)\sum_{z \in Z} P(w|z)P(z|d) \quad (2)$$

Using the frequency of the words in documents ($n(d,w)$) it is possible to rearrange the probabilities to develop an iterative expectation maximization scheme to estimate all the

probabilities. The expectation step generates P(z|d,w) while the maximization step calculates P(w|z), P(d|z) and P(z). The distances in semantic feature space can be plotted from the PCA (i.e. the axes do not relate to a specific attribute) and the distances between classes can be visualised in this PCA space. The descriptions of the land cover classes were from "Guidelines for the selection of Local Wildlife Sites in Leicester, Leicestershire and Rutland" and the LCM2007 descriptions were from http://www.ceh.ac.uk/sci_programmes/BioGeoChem/DatasetInformation.html

IV. RESULTS

A. Data Primitives

The degree of overlap in species richness between different classes are calculated in each dimension or domain. Note that the matrix of overlaps may not be symmetrical – i.e. that $O(p_A, p_B) \neq O(p_B, p_A)$. Overlaps from different dimensions are averaged and then used to generate two look up tables: *National to Local* and *Local to National* as shown in Table 2.

The tables can be used to infer the presence of one type of habitat (e.g. local) from the other (e.g. national) for the LWRT data and for the LCM2007 data. The degree of overlap between the different types of habitat are shown in Figure 1 for the local habitat of Ancient Woodland, inferred from the LCM2007 data for the Benscliffe wood area.

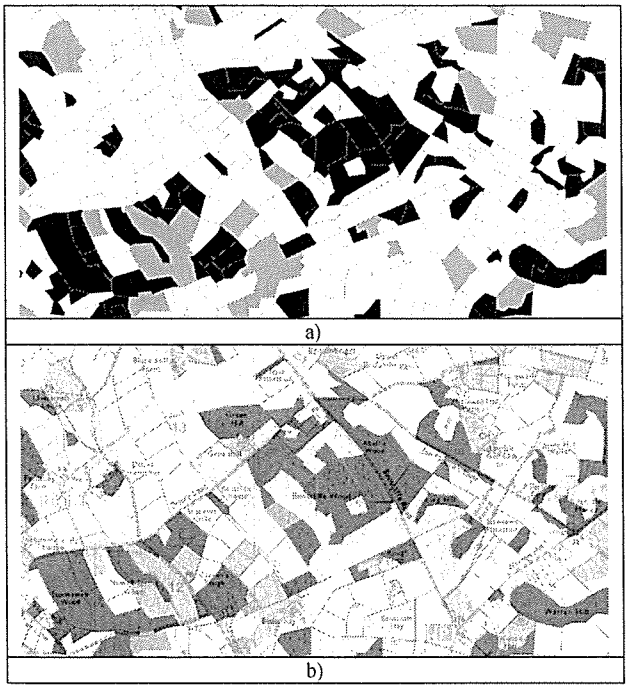

Figure 1. a) Areas of 'Ancient Woodland' inferred from LCM2007, and b) with context from an OS basemap with their shade corresponding to the 'belief' in that area. (© Crown Copyright/database right 2010. An Ordnance Survey/EDINA supplied service).

B. Latent Semantic Analysis

The PLSA identifed 12 dimensions or topics from an anlysis of the full text of the 15 local habitat classes and the 17 national broad. Figure 2 shows the concepts associated with each of the PLSA topics and Figure 3 shows how each the links between these and the different classes from the two classification schema.

TABLE II. AVERAGE OVERLPAS BETWEEN NATIONAL AND LOCAL HABITATS

From National	Acid Grass	Mixed Grass	Broadleaved Woodland	From Local	Improved Grassland	Neutral Grassland	Broadleaved, Mixed and Yew woodland
Improved Grassland	0.42	0.25	0.19	Acid Grass	0.36	0.64	0.31
Neutral Grassland	0.51	0.82	0.25	Mixed Grass	0.25	0.9	0.52
Broad-leaved, Mixed and Yew woodland	0.12	0.32	0.72	Broad-leaved Woodland	0.14	0.31	0.83

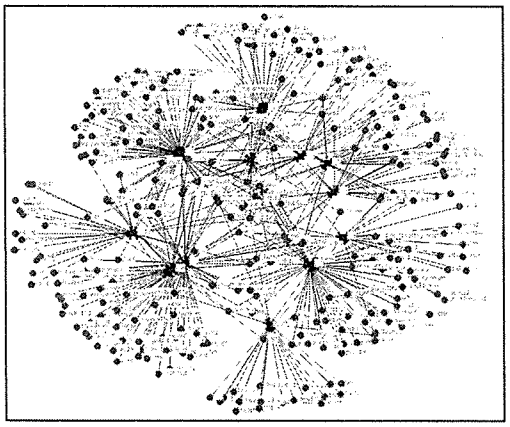

Figure 2. The topics identified by PLSA (blue) and the terms associated with each of these

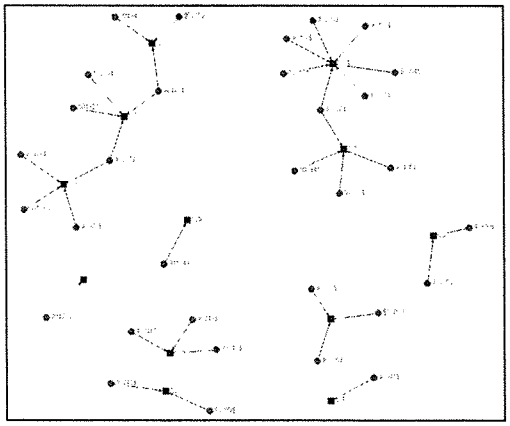

Figure 3. The classes (red) associated with each of the PLSA terms (blue).

The terms associated with each dimension provide insight into the semantic concepts with which it is associated. Table 3 shows the terms most strongly associated with 2

dimensions: 11 relates to woodland and 10 to bare ground habitats.

TABLE III. CONCEPTS ASSOCIATED WITH 2 EXAMPLE LATENT SEMANTIC DIMENSIONS

11	10
broadleaved	aira
canopy	bare
commercial	birds
coniferous	built
covers	category
gardens	diversity
total	fern
urban	ground
woody	hair-grass
	insects
	lirbap
	parmelia
	rocks
	structures
	ukbap

V. DISCUSSION AND CONCLUSION

The results of applying the two approaches are not directly comparable. The data primitive approach assumes that the important dimensions are known to the user. The application of latent semantic analysis 'reveals' the important dimensions which may be unknown to the user. Both approaches can be considered bottom up, in that the user specifies what is important to them in the integrating activity. Additionally, the final, and perhaps most difficult, task of bridging the top-down and bottom-up approaches has yet to be attempted within both formal ontology research activities such as OWL and emerging e-science infrastructures such as INSPIRE. Both of these approaches provide methods to support activities that seek to encode data semantics in reference frameworks, whether those dimensions are known a priori (data primitives) or not (latent semantic analysis).

REFERENCES

Ahlqvist, O. (2004). A parameterized representation of uncertain conceptual spaces. *Transactions in GIS*. 8, 493–514.

Ahlqvist, O. (2008). In search of classification that supports the dynamics of science: the FAO Land Cover Classification System and proposed modifications. *Environment and Planning B: Planning and Design*. 35, 169-186.

Blei, D.M., Ng, A.Y. Jordan M.I. (2003). Latent dirichlet allocation. *Journal of machine Learning Research*. 3, 993-1022.

Bouchon-Meunier, B., Rifqi, M. and Bothorel, S. (1996). Towards general measures of comparison objects. *Fuzzy sets and systems*. 84, 143-153.

Comber, A.J., Fisher, P.F. and Wadsworth, R.A., (2008a). Semantics, metadata, geographical information and users. *Transactions in GIS*. 12(3), 287–291.

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (2008b). Using semantics to clarify the conceptual confusion between land cover and land use: the example of 'forest'. *Journal of Land Use Science*. 3(2-3), 185-198.

Comber, A.J., Fisher, P.F., Wadsworth, R.A., (2004a). Assessment of a semantic statistical approach to detecting land cover change using inconsistent data sets. *Photogrammetric Engineering and Remote Sensing*. 70(8), 931-938.

Comber, A., Fisher, P., Wadsworth, R., (2004b). Integrating land cover data with different ontologies: identifying change from inconsistency. *International Journal of Geographical Information Science*. 18(7), 691-708.

Comber, A., (2008). The separation of land cover from land use with data primitives. *Journal of Land use Science*. 3(4), 215–229.

Di Gregorio, A., and Jansen, L.J.M. (2000). *Land cover classification system: classification concepts and user manual*, Rome: FAO.

Frank A.U., (2007a). Towards a mathematical theory for snapshot and temporal formal ontologies Pages 317-334 in *The european information society, lecture notes in geoinformation and cartography*, Springer Berlin Heidelberg

Frank, A.U., (2001). Tiers of ontology and consistency constraints in geographical information systems, *International Journal of Geographical Information Science*. 15 (7), 667-678.

Gallagher, P. and Calder, G. (2007). *Biodiversity of scottish wildlife trust reserves* http://www.swt.org.uk/docs/002_001_publications_BiodiversitySWTReserves07_1250 595197.pdf [Available 17/08/09]

Harvey F, Kuhn W, Pundt H, Bishr Y, Riedemann C (1999) Semantic interoperability: A central issue for sharing geographic information. *Annals of Regional Science*. 33(2), 213-232.

Hofmann, T (1999a). Probabilistic latent semantic indexing, in Hearst M, Gey F, Tong R (Eds) *Proceedings of 22nd international conference on research and development in information retrieval*. (pp 50-57). Univ Ca, Berkeley, California, Aug, 1999.

Honkela T., (1997). *Self-organising maps in natural language processing*. PhD thesis Helsinki University of Technology, Department of Computer Science and Engineering, http://www.cis.hut.fi/~tho/thesis/.

Kavouras M. and Kokla M. (2002). A method for the formalization and integration of geographical categorizations. *International Journal of Geographical Information Science*. 16: 439-453.

Lin, X., (1997). Map displays for information retrieval. *Journal of the American Society for Information Science*. 48:40-54.

NE and TWT (Natural England and the Wildlife Trusts) (2009). *6Cs Growth point biodiversity opportunity mapping*. http://www.emgin.co.uk/default.asp?PageID=261_[available 17/08/09].

Pundt, H. and Y. Bishr, (2002). Domain ontologies for data sharing-an example from environmental monitoring using field GIS. *Computers and Geosciences*. 28 (1): 95-102.

Wadsworth R.A, Comber A.J., and Fisher P.F., (2006). Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. in Andreas Riedl, Wolfgang Kainz, Gregory Elmes (eds.). *Progress in spatial data handling, proceedings of SDH 2006*, (pp 197–213). Berlin: Springer.

Wadsworth R.A., Fisher P.F., Comber A., George C., Gerard F. and Baltzer H. (2005). Use of quantified conceptual overlaps to reconcile inconsistent data sets. Session 13 Conceptual and cognitive representation. *Proceedings of GIS planet 2005*, Estoril Portugal 30th May - 2nd June 2005. ISBN 972-97367-5-8. 13pp

Wadsworth, R., Balzter, H., Gerrard, F., George, C., Comber, A.J., and Fisher, P.F. (2008b). An environmental assessment of land cover and land use change in central siberia using quantified conceptual overlaps to reconcile inconsistent data sets. *Journal of Land Use Science*, 3(4): 251.

Wadsworth, R.A, Comber, A.J. and Fisher, P.F., (2008a). Probabilistic Latent Semantic Analysis as a potential method for integrating spatial data concepts, pp 99-108 in Gerhard Navratil (ed). *Proceedings of the colloquium for Andrew U. Frank's 60th Birthday*, GeoInfo Series 39, Vienna.