Thoughts on exploiting instability in lattices for assessing the discrimination adequacy of a taxonomy

Antony K Cooper⋆, Derrick G Kourie, and Serena Coetzee

Department of Computer Science, University of Pretoria, Pretoria, South Africa
acooperatcsir.co.za,dkourieatcs.up.ac.za,scoetzeeatcs.up.ac.za
http://www.cs.up.ac.za/

Abstract. Conventionally in formal concept analysis (FCA), concept stability is preferred in the lattice, because instability (i.e. low stability) represents noise that clouds the analysis of the data.
However, high concept stability means that there are many objects with the same intent and many attributes with the same extent, which could be interpreted as a high level of redundancy in the lattice. We report here on work that we have done using FCA to analyse different taxonomies for user-generated content. Here, redundancy amongst the attributes in the lattice is not desirable, because it represents classes in the taxonomy that are unable to differentiate adequately the objects being classified from one another. Low extensional stability (i.e. noisy attributes) reveals attributes that are unique to the associated set of objects — if the attributes are interesting, they could imply there are "missing" objects. Redundancy amongst the objects can have a number of implications. Hence, instability in a lattice is desirable for some types of analysis.

Keywords: formal concept analysis, stability, discrimination adequacy, taxonomy

1 Background on user generated content

User-generated content (UGC) in general, and *volunteered geographical information (VGI)* in particular, are becoming more important as sources for official data bases, such as those used in national *spatial data infrastructures (SDIs)*. One of the distinguishing characteristics of the use of spatial data is that the same, common, base data sets are used by many different users for many diverse applications. Hence, there is a growing need to share and organise spatial data across different disciplines and organisations, which has resulted in the development and implementation of SDIs and of the theory and notions behind them. An SDI is an evolving concept about facilitating and coordinating the exchange and sharing of spatial data and services between stakeholders from different levels in the spatial data community [1].

⋆ Corresponding author. Current address: Built Environment Unit, CSIR, PO Box 395, Pretoria, 0001, South Africa

However, while the traditional sources of official data are well understood, the same does not apply to VGI nor other forms of UGC. These concepts are interpreted in different ways, and one woman's user generated content could be another man's professionally generated content. Several attempts have been made to understand such data and the contributors of such data, by developing taxonomies for aspects of UGC in general (eg: [2, 3]), or of VGI in particular (eg: [4, 5]). Examples of VGI include *OpenStreetMap*, a free, editable map of the whole world [6]; citizen-science projects such as the Second South African Bird Atlas Project (SABAP2) [8]; in-car navigation systems allowing users to submit corrections and updates to the map data; and geocoded photographs and Wikipedia [7] entries contributed to virtual globes such as *Google Earth* [9].

We have conducted an assessment amongst some geographical information professionals of their perceptions of virtual globes, VGI and SDIs [10], and we are in the process of developing a taxonomy of VGI, which we are modelling formally. We have used *formal concept analysis (FCA)* [11] to assess the *discrimination adequacy* and other characteristics of these existing taxonomies, rather than to classify UGC or VGI. Specifically, we are using FCA to assess if there are repositories of UGC that are not classified by a taxonomy; classes in a taxonomy for which there are few or no repositories; or classes that are not differentiated from one another. The intention is to improve the understanding of UGC and VGI, such as for assessing the quality of any set of VGI, and for catering for VGI in a standard such as ISO 19115:2003, *Geographic information — Metadata*.

2 Background on formal concept analysis

Formal concept analysis (FCA) essentially uses a lattice of formal concepts and formal attributes and the linkages between them, for data analysis, knowledge representation and information management. We use here the standard terminology and notation for FCA, see [11–16].

A *lattice* is an *ordered set*, denoted as $(P; \leq)$, where for any pair of elements x and y in P, both the *supremum* (the least upper bound, or the *join*), $x \vee y$, and the *infimum* (the greatest lower bound, or the *meet*), $x \wedge y$, always exist. Such an ordered set is a *complete lattice* if the supremum, $\bigvee S$, and the infimum, $\bigwedge S$, exist for any subset S of P. A complete lattice always has a top element (the *unit*), that is greater than all the other elements, and the dual of this, a bottom element (the *zero*), that is smaller than all the other elements. In an ordered set, element x is *covered* by y if $x < y$ and there is no $z \in P$ such that $x < z < y$. This is denoted as $x \prec y$.

A *formal context* is written as: $\mathbb{K} := (G, M, I)$, where G is a set of *formal objects* (known as the *extent* of the concept), and M is a set of *formal attributes* known as the *intent* of the concept), such that M contains all the attributes that the objects in G have in common, and only those attributes. I is the binary relation between the sets of objects and attributes, namely: $I \subseteq (G \times M)$.

This is a *one-valued context*, in that each attribute has one of only two values, *present* or *absent*. In a *many-valued context*, attributes can have a domain of multiple values (eg: a date or a count). A *many-valued context* is a quadruple, $\mathbb{K} := (G, M, V, I)$, where G is a set of objects, M a set of many-valued attributes, V a set of attribute values (from a domain) and I the ternary relation, $I \subseteq (G \times M \times V)$, such that: $(g, m, v) \in I$ and $(g, m, w) \in I$ always implies that $v = w$.

Any many-valued context can be transformed to a one-valued context by replacing every valid attribute and attribute value pair by a new attribute, in a process is known as *conceptual scaling*. For the analysis discussed here, one-valued contexts have been used because the lattices are clearer, even though they have more attributes.

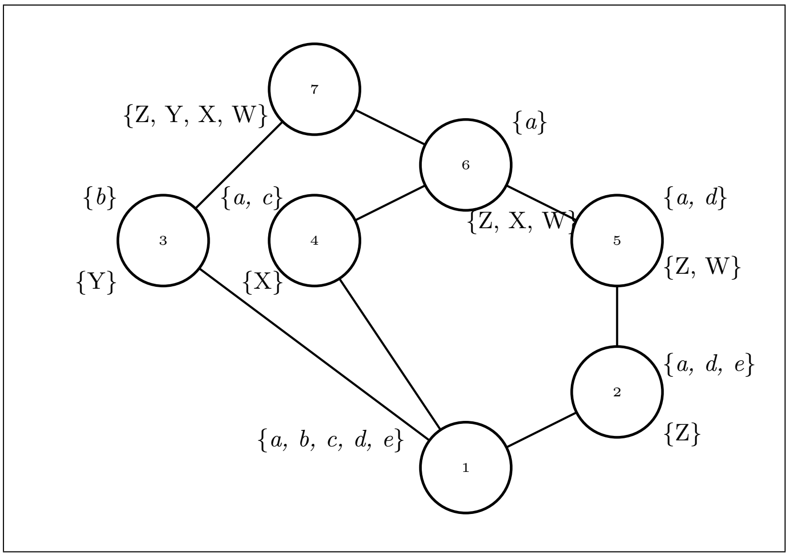


Fig. 1. An example of a line diagram.

Every finite ordered set $(P; \leq)$ can be drawn and if $x \prec y$ and if element x is placed below element y in the diagram, then it is known as a *line diagram* or a *Hasse diagram*. Figure 1 shows a line diagram of a complete lattice, where node 1 is the zero of the lattice, and node 7 the unit. Node 5 covers node 2 and node 2 is covered by node 5, for example. Attached to each node are objects (the upper-case letters) and attributes (the lower-case letters).

A formal context can also be considered as a table, relating objects to attributes — indeed, FCA tools such as ConExp (Concept Explorer) [17] and ConImp (Contexts and Implications) [18] use a table for inputing the lattice. For example, Table 1 shows the formal context in Figure 1, where a cross in cell ij indicates that object i (in row i) is described by attribute j (in column j)).

Table 1. A cross-table of the formal context shown in Figure 1.

	a	b	c	d	e
Z	×			×	×
Y		×			
X	×		×		
W	×			×	

3 Stability in a lattice

The *stability* of a formal concept is an indication of how much its intent depends on individual objects in the extent, and how much the extent depends on individual attributes in the intent. Thus, the intensional stability of a concept is a measure of the likelihood that removing a random set of objects from the concept's extent would change its intent. Similarly, extensional stability measures the likelihood that removing a random set of attributes from a concept's intent would change its extent.

Formally, the *intensional stability index*, σ_i, and the *extensional stability index*, σ_e, of concept (A, B), are defined in [16] as follows:

$$\sigma_i(A,B) = \frac{\mid \{C \subseteq A \mid C' = B\} \mid}{2^{|A|}}$$

$$\sigma_e(A,B) = \frac{\mid \{D \subseteq B \mid D' = A\} \mid}{2^{|B|}}$$

To explain the intuition underlying these definitions, note that each concept (A, B) has $|A|$ objects in its extent. Clearly, a total of $2^{|A|}$ subsets of such objects can be formed. Now suppose that, instead of the objects in A, an arbitrary one of these subsets, say $C \subseteq A$, had been used to construct the lattice whose context had otherwise remained unchanged. Such a lattice would then contain a concept (C, C'), and the intent of this concept would be related to the intent of the original concept by $C' \supseteq B$. Now the intentional stability of the original concept, $\sigma_i(A, B)$, is measured by the proportion of object subsets in the given context, such as C, that have the specific property that $C' = B$. The intent of concept (A, B) is therefore "stable" whenever any one (or more) of these subsets of objects such as C is used to construct a lattice, assuming that the rest of the context remains unchanged. Conversely, if a new lattice is built whose context does not have any single such subset of objects, then the resulting lattice will no longer have a concept whose intent is B. Formally, when $\sigma_i(A, B) = 0$, each and every object in these subsets has at least one attribute that is not in the intent of (A, B).

The notion of extensional stability is similar, but with the roles of extents and intents reversed: the extent of concept (A, B) is stable with respect to an intent subset, D, if A is retained as the extent of a concept in a revised lattice whose attributes are D rather than B. The number of such subsets, D, relative to the total number of possible subsets of B provides a stability measure for the concept. Formally, when $\sigma_e(A, B) = 1$, all the subsets of attributes of (A, B) have the same extent as (A, B), and when $\sigma_e(A, B) = 0$, each and every attribute in these subsets has at least one object that is not in the extent of (A, B).

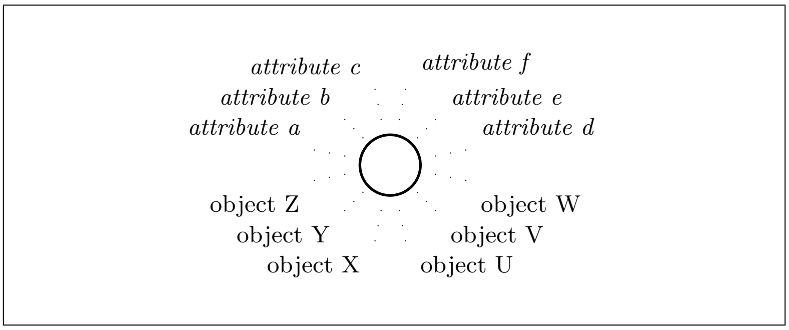


Fig. 2. A very stable, but rather boring, lattice.

Unsurprisingly, the more objects (or attributes) covered by a formal concept, the more likely it will be intensionally stable (or extensionally stable), because of the greater likelihood there will be "redundant" objects (or attributes). Figure 2 shows a lattice that is very stable, because all the attributes are in the intent of each and every object, and all the objects are in the extent of each and every attribute. That is, $\sigma_i(A, B) = 1$ and $\sigma_e(A, B) = 1$. However, while this is an extreme example of stability, it does illustrate why stability can mean redundancy and why stability can be considered "boring" in some applications [19], because of the low information content.

We appreciate that when FCA is being used for machine learning (eg: [19]), if concepts in the resulting lattice have high stability it means that the input data were robust with little noise (eg: caused by coding errors or instrument accuracy). Instability (i.e. low stability) represents noise that clouds such analysis of the data.

However, as outlined in Section 1, we have not used FCA to classify data, but to assess the *discrimination adequacy* of the taxonomies for user-generated content of [2, 3, 5, 4]. The classes in these taxonomies are the attributes for FCA. For example, [3] provides four classes for copyright issues to extend the taxonomy in [2], namely *User-authored content*, *User-derived content*, *User-copied content* and *Peer-to-peer as UGC*. For the analysis reported on here, the objects are repositories of user-generated content (not necessarily available on the Internet). [5] provides examples of such repositories (eg: in-car navigation, or an open repository), which we have supplemented, based on our experience (eg: traditional spatial data infrastructure (SDI) with strict control over its data sources, or revision requests submitted to an SDI). The assignment of attributes to objects is based on the discussions of their taxonomies by [2, 3, 5, 4], and our judgement.

For our analysis then, formal concepts with low stability are generally more interesting, because of their unique objects and/or attributes — it is more appropriate to consider these to be extreme values, rather than noise. We are using FCA to examine pre-existing taxonomies to determine their *discrimination adequacy* in classifying a target set of repositories of user-generated content, that is:

- Are there isolated formal objects with few or no formal attributes in their intent? This would indicate repositories that are not classified by the taxonomy, and hence classes (formal attributes) that are missing.
- Are there isolated formal attributes with few or no formal objects in their extent? This would indicate classes in the taxonomy for which there are few or no repositories, and hence repositories (formal objects) that are missing.
- Are there attributes that are redundant? This would indicate taxonomy classes that are not differentiated from one another, which could be because they are redundant or poorly defined.

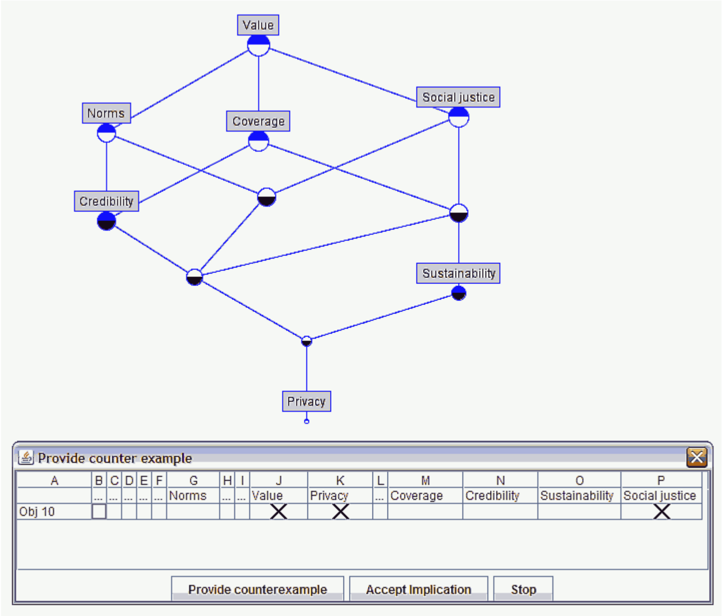
The implications of these are discussed below in Section 5. In our analysis, we expect to get redundant objects, because in the real world, there are likely to be such repositories with identical classifications.

4 Attribute exploration

The tool used to support FCA for this analysis is Concept Explorer (ConExp) [17]. It was selected because it is an open-source tool, is robust and is used by other researchers in the Department of Computer Science at the University of Pretoria (see, for example, [19–22]), and hence has a pool of expertise that is readily available to us.

For our analysis discussed here, the key functionality provided by ConExp is *attribute exploration*. This is an interactive process to see if each *implication* (sets of "linked" formal attributes) can also apply to formal objects that are not in the context of the implication. Questions are asked about dependencies between different formal attributes from some fixed set of formal attributes (i.e. the exploration), and if a dependency does not hold, the user has to provide a counterexample (effectively, the user must add a new formal object) [17]. Attribute exploration can then reveal "missing" attributes, "missing" objects, "redundant" attributes and "redundant" objects.

Figure 3 shows a screen shot of ConExp doing attribute exploration. Here, we show a subset of the classes from the taxonomy of [4], and one can see that the attribute *Privacy* is attached to the zero of the lattice — in other words, none of the objects (i.e. repositories) have *Privacy* as an attribute. In doing this attribute exploration, the user was asked to provide a counter example of an object with *Privacy* and *Value* as attributes, with the default name of *Obj 10*, and this the user was doing, while specifying an additional attribute, *Social*



Provide counter example

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	...	...	...	...	...	Norms	...	...	Value	Privacy	...	Coverage	Credibility	Sustainability	Social justice
Obj 10	□								X	X					X

Fig. 3. Attribute exploration in ConExp.

Justice, for the object. Please note that the names of the repositories have been omitted for clarity, and because this was not meant to be a definitive analysis of the repositories.

In terms of the duality principle, the formal objects and the formal attributes can be swopped in such a tool for analysis (i.e. the axes of the matrix transposed), if such analysis would be useful to explore. In practice, this would put the formal objects (i.e. the repositories in our case) into the intent of the concept and the formal attributes (i.e. the taxonomy classes) into the extent. Effectively, this would allow one to do *object exploration*. The benefit of doing this in ConExp is that one can then determine the similarity of different objects, for example.

5 Missing and redundant attributes and objects

Formal concept analysis is applied here to determine the adequacy of these taxonomies for discriminating between selected repositories on the Web containing *user generated content (UGC)* in general, or *volunteered geographical information* in particular. Surprisingly, there appears to have been few attempts yet at developing a taxonomy of user generated content, with the most comprehensive having been compiled by [5], with a specific focus on VGI. Other such taxonomies are [2–4].

As discussed in Section 2, a formal context can be *one-valued* or *many-valued*. *These taxonomies in [2–4] are generally* many-valued contexts*, but often without* *the domains of valid values being specified. It can be confusing and meaningless to use such attributes without their values in FCA.* While an attribute such as *user-authored content* could be interpreted by the reader as meaning that the objects (UGC repositories) in its extent were authored by the user (contributor), for example, an attribute such as *distribution platform* cannot be so interpreted and having the attribute *quality* does not say if the quality is good or bad. Hence, it has been necessary to add attribute values in several instances for the analysis. Hopefully, these additions are appropriate.

Through attribute analysis, FCA can identify "missing" and "redundant" attributes and objects, as discussed below.

5.1 Missing formal attribute

A missing formal attribute would occur when there are two or more objects in an extent that one would expect to be differentiated from one another by their attributes, but they are not. The extensional stability would be low and FCA would highlight where this attribute is missing. The problem could be addressed by defining one or more suitable attributes for the intent of these objects, to separate them. In the context of our analysis, this would involve adding one or more classes to a taxonomy, so that the taxonomy would differentiate better between the repositories. In other words, the taxonomy would be improved by adding the class.

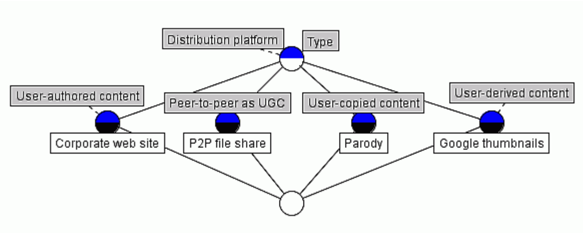


Fig. 4. The taxonomy of [3] for copyright issues for UGC.

The taxonomy in [3] is an extension of the taxonomy in [2], to enable the latter to cater for copyright issues. Figure 4 shows the four classes for copyright issues added to the two main axes of the taxonomy in [2], which were *Distribution platform* and *Type*. As can be seen, removing the four classes for copyright issues would collapse the lattice into a very stable form, such as is shown in Figure 2

5.2 Redundant formal attribute

Where redundant formal attributes occur, this could be coincidental, could reflect a set of objects that is too narrow (eg: there are other types of repositories that should have been included in the analysis), or could indicate that some

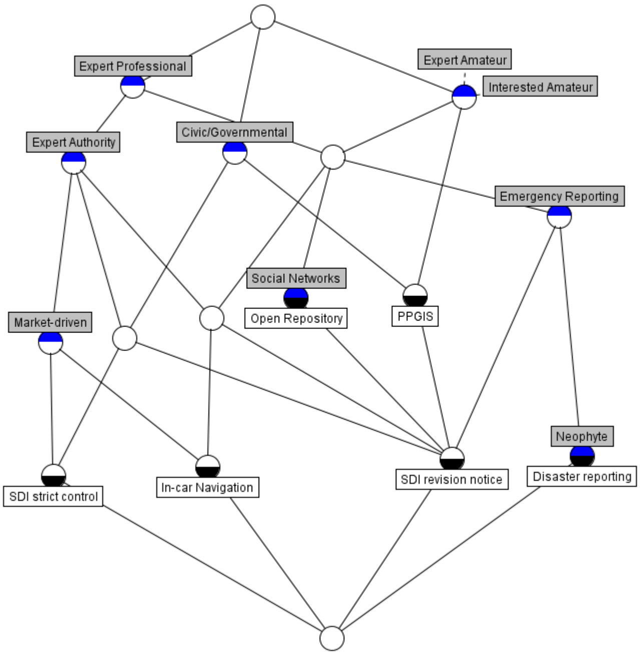


Fig. 5. A subset of the taxonomy of [5] for assessing the nature and motivation of *produsers*.

classes should be removed from the taxonomy because they add no value or even worse, could cause confusion as users try to differentiate between classes that are, in essence, equivalent. From the FCA perspective, the intensional stability is too high.

We illustrate this with a subset of the taxonomy developed by [5], for assessing the nature and motivation of *produsers* (that is, users who are also producers). For the objects, we use the generic examples given by [5] in their Table 1, namely in-car navigation (eg: Tom Tom, Tele Atlas or NAVTEQ), open repository (eg: OpenStreetMap), public participatory geographical information system (PPGIS) and disaster reporting (eg: during the recent Haiti earthquake). As this set is a bit limited, we add to this a traditional SDI with strict control over its data sources, and revision requests or notices submitted to an SDI (eg: as described by [23] for swisstopo).

Figure 5 shows the line diagram of this concept lattice. As can be seen, in this set of objects (spatial data repositories) and attributes (taxonomy classes), there is redundancy in the attributes *Interested Amateur* and *Expert Amateur*. This means that these two classes are inadequately defined, or cannot be differentiated in practice, or other types of repositories should be included in this analysis. In this case, the problem appears to be with differentiating between the classes in practice, because both interested and expert amateurs are likely to make the same types of contributions of UGC to the same types of repositories. Further, Figure 6 shows the lattice in Figure 5, but with just the repositories given in Table 1 of [5]. One can see clearly the increase in the redundancy of the attributes.

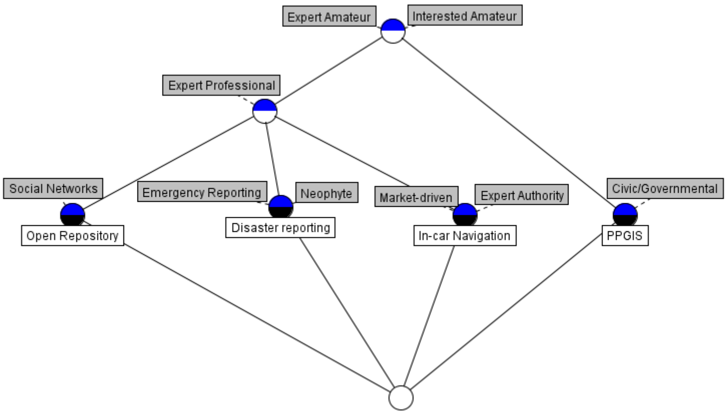


Fig. 6. Figure 5 with only the repositories in [5].

5.3 Missing formal object

A missing formal object would be a type of repository that has not been included in the analysis. This could be a weakness in the analysis, in that an important type of repository had been omitted. Alternatively, it could indicate a type of repository that does not yet exist and hence a potential "gap" in the market — revealed because of the low intensional stability.

It was while experimenting with FCA and the taxonomy of [4], that we first discovered the value of instability in a lattice. This is to some extent an artificial example as it was not meant to be a definitive analysis of the repositories. However, as shown in Figure 3, it does illustrate a potential "gap" in the market — in this case, it would appear that repositories do not cater adequately for privacy, a widespread problem on the Internet.

5.4 Redundant formal object

In a comprehensive analysis of repositories of UGC, one would expect to find redundant objects, that is, repositories that are fundamentally equivalent and hence direct competitors of one another, though possibly targeting different domains (assuming that the taxonomies do not differentiate on the domains). For example, referring to Figure 5, the objects are generic or abstract, and there could be many repositories that are specific instances of each. Adding these repositories as objects would create redundancies. From the FCA perspective, this would be high extensional stability.

6 Conclusions

We have used formal concept analysis (FCA) to assess the adequacy of several taxonomies of user-generated content (UGC) in discriminating between different types of repositories of UGC. In contrast to the usual applications of FCA, we have shown that instability (i.e. low stability) in a lattice can have value for analysis. For our analysis, high intensional stability reveals taxonomy classes that are redundant; high extensional stability reveals redundancy amongst the repositories, which is to be expected; low intentional stability reveals missing repositories or gaps in the market; while low extensional stability reveals missing classes from the taxonomy. Hence, instability in a lattice is useful for some types of analysis.

We have reported here on preliminary work that we have done to assess several taxonomies of UGC. Future work could involve expanding the analysis to assess each taxonomy *in toto*, or a combination of the taxonomies. We are using our analysis to inform our own work on developing a taxonomy of volunteered geographical information, and we could also use FCA to assess the discrimination adequacy of this taxonomy. We could also apply FCA to assess taxonomies in other domains, such as to build on previous work we have done on bloodstain pattern analysis [24]. Finally, as we alluded to in Section 4, we could explore the utility of *object exploration*.

We would like to acknowledge the fruitful discussions we have had with our colleagues to understand better the theory and applications of FCA.

References

1. Hjelmager, J., Moellering, H., Delgado, T., Cooper, A.K., Rajabifard, A., Rapant, P., Danko, D., Huet, M., Laurent, D., Aalders, H.J.G.L., Iwaniak, A., Abad, P., Düren, U., Martynenko, A.: An initial Formal Model for Spatial Data Infrastructures. International Journal of Geographical Information Science, 22(11), pp 1295–1309 (2008)
2. Wunsch-Vincent, S., Vickery, G.: Participative Web: User-Created Content. Organization for Economic Co-operation and Development, report number DSTI/ICCP/IE(2006)7/FINAL. Compiled for the Working Party on the Information Economy of the Committee for Information, Computer and Communications Policy (2007)

3. Gervais, D.: The Tangled Web of UGC: Making Copyright Sense of User-Generated Content. Vanderbilt Journal of Entertainment and Technology Law, 11(4), pp 841–870 (2009)
4. Budhathoki, N.R., Nedovic-Budic, Z., Bruce, B.: An interdisciplinary frame for understanding volunteered geographic information. Geomatica, 64(1), pp 11–26 (2010)
5. Coleman, D.J., Georgiadou, Y., Labonte, J.: Volunteered Geographic Information: The Nature and Motivation of Produsers. International Journal of Spatial Data Infrastructures Research, Special Issue on GSDI-11, 4 (2009)
6. OpenStreetMap: The Free Wiki World Map. Home page. Accessed 13 June 2010 at: http://www.openstreetmap.org/ (2010)
7. Wikipedia. Home page. Accessed 13 June 2010 at: URL http://en.wikipedia.org/ (2010)
8. Animal Demography Unit: Southern African Bird Atlas Project 2. Home page. Accessed 13 June 2010 at: http://sabap2.adu.org.za/ (2010)
9. Google Earth: Explore, Search, and Discover. Home page. Accessed 13 June 2010 at: http://earth.google.com/ (2010)
10. Cooper, A.K., Coetzee, S., Kourie, D.G.: Perceptions of virtual globes, volunteered geographical information and spatial data infrastructures. Geomatica, 64(1), pp 333348 (2010) ctures. Geomatica, 64(1), pp 333–348 (2010)
11. Wille, R.: Restructuring lattice theory: An approach based on hierarchies of concepts. In: Ordered sets, Rival, I. (ed), pp 445–470, D Reidel Publishing Company, Dordrecht-Boston (1982)
12. Ganter, B., Wille, R.: Applied Lattice Theory: Formal Concept Analysis. Preprints. 14pp (1997)
13. Carpineto, C., Romano, G.: Concept Data Analysis: Theory and Applications. John Wiley & Sons, Ltd (2004)
14. Priss, U.: Formal concept analysis in information science. Annual review of information science and technology, 40, pp 521–543 (2006)
15. Kuznetsov, S.O.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence, 49(1–4), pp 101–115 (2007)
16. Klimushkin, M., Obiedkov, S., Roth, C.: Approaches to the Selection of Relevant Concepts in the Case of Noisy Data. In: 8th International Conference on Formal Concept Analysis (ICFCA 2010), Agadir, Morocco, pp 255–266, Springer (2010)
17. Yevtushenko, S., Kaiser, T., Tane, J.: Concept Explorer The User Guide. (2003)
18. Burmeister, P.: Formal Concept Analysis with ConImp: Introduction to the Basic Features. (2003)
19. Kourie, D.G., Oosthuizen, G.D.: Lattices in machine learning: Complexity issues. Acta Informatica, 35, pp 269–292 (1998)
20. Cleophas, L., Watson, B.W., Kourie, D.G., Boake, A., Obiedkov, S.: TABASCO: a Taxonomy-based Domain Engineering Method. South African Computer Journal, (37), pp 30–40 (2006)
21. Obiedkov, S., Kourie, D.G., Eloff, J.H.P.: Building Access Control Models with Attribute Exploration. Computers and Security, 28(1–2), pp 2–7 (2009)
22. Chan, K.S.M.: Formal Methods for Web Services: A Taxonomic Approach. In: 32nd International Conference on Software Engineering (ICSE'10), Cape Town, South Africa, 2, pp 357–360, ACM (2010)
23. Gulat, J.-C.: Integration of user generated content into national databases — Revision workflow at swisstopo. 1st EuroSDR Workshop on Crowd Sourcing for Updating National Databases, Wabern, Switzerland (2009)
24. Cooper, A.K. Thoughts on categorising bloodstain patterns. Technical Report 0442-0001-701-A1, CSIR (2003)