Paper: A discretization-hierarchization approach to information management -- implications for Digital Earth

Abstract: The paper addresses ways of supporting the Digital Earth project. Highlighting the merits of discretization and analyzing classification-hierarchization operations as forms of discretization, it analyzes also their weaknesses and shows that their rigidity and inertia stand in contrast to the flexibility required by Digital Earth, and proposes a path towards solutions.
Résumé:

1. Aims and scope of the paper. Digital Earth.

This paper is not about library science. However, due to the general nature of this investigation, links to library science cannot be excluded. It is not about Google Earth, World Wind, or other geobrowsers either. Its exploration is not confined to geographical information systems. The paper focuses on some critical aspects of a vast knowledge-managing project, Digital Earth (DE); it identifies discretization and hierarchization as crucial information-handling processes, and explores their implications for DE effectiveness. Discretization and hierarchization, as well as their implementation through classification, are applied in DE in a very wide variety of situations: rather than approaching them in concrete instances and proceeding to generalizations, this study is performed at the conceptual level.

Emerged from a pragmatic initiative launched at the end of the last millennium (Gore 1999), the DE project includes a fast growing, increasingly integrated set of bodies of knowledge, based on collaboration from the local to the global level and focusing on geo-referenced information synergically handled to offer effective access to a variety of users, from students to governments and experts in the industry and the Academia (Craglia et al. 2012). It represents more than technology, methods, and standards: it announces a novel way of addressing information management (Suteanu 2011).

The information to be included and integrated in the DE system is extremely diverse: it extends from spatial information on maps and satellite images to time series collected from space or on the ground, to bodies of knowledge from the most different disciplines (for instance, historical and anthropological aspects of certain themes of interest, information about chemical reactions in the atmosphere or in the oceans, taxonomical information regarding organisms from a certain region), to outcomes of laboratory experiments, results of social science studies (e.g. local perception of natural hazards) etc. The amount of collected information and its diversity are both growing fast. The tasks of integration become more and more challenging. For example, a study regarding renewable energy should be able to draw from DE information about: wind pattern statistics on different scales (from minutes to hours to months) in different locations, the electric grids in the region, the spatial distribution and consumption patterns of electricity users, engineering challenges involved in the development of the necessary grids, local attitude towards renewable energy, potential costs for different possible projects, potential industry partners, policies and the legislative framework, projections regarding

future patterns – from energy consumption to technological progress to costs etc.; moreover, the DE system should be able to continuously and smoothly absorb changes to already existing aspects of the study, new challenges, new solutions etc. The sources for these informational components are not created for the project itself: a key strength of DE consists of its capability of using information that was provided by innumerable users working in different disciplines, and integrating it in a way that makes it applicable to many potential applications, some of which are not even foreseen when information is collected.

The fact that in this case information is (i) not specifically collected and prepared from the beginning for a certain purpose, and (ii) characterized by a large variety of acquisition conditions and formats, starting from different premises, based on different ways of structuring and representing information etc., makes integration both important and difficult. The heterogeneity that characterizes this material is the source of some of the greatest challenges (Parsons *et al*. 2011). In this context, discretization and hirerachization are expected to play a key role for the outcomes of the DE system.

2. Information flux and discretization

Since a key feature of information is related to novelty and transformation, one can distinguish "information flux" or "i-flux" as an aspect of information defined as the specification of processes, or "change" (Suteanu 2010b). "Change" may refer to different moments in time, different portions of space, or even different scales (Suteanu 2010a). In this context, a transition between similar configurations involves a lower information flux density than one involving states that are very different from each other. The i-flux that corresponds to two states that are compared can be called a state quantity, as opposed to a process quantity that would depend on the path followed by the considered system from one state to the other. I-flux is often important to information management, since it may significantly influence the resources required for information handling in terms of time, memory, transmission capacity etc.

This paper argues that among the operations performed upon the collected information, discretization and classification are critical for the resulting effectiveness of the DE system. Both are addressed here in the most general sense, meant to include a wide diversity of situations encountered in DE (examples can be as different as classification regarding slopes on a map, categories of fluctuations in river discharge, categories of myths of a given population, types of chemical reactions, plant taxonomy, etc.).

It has been shown that analogue features cannot be identical, and only discrete configurations, characterized by a finite number of states, may lead to identicalness and thereby potentially to zero i-flux (Suteanu 2007). Whether we are facing an essentially discrete nature of the material world or not (Fredkin 2004) is thus crucial in terms of our understanding of i-flux, no matter how small the actual differences between distinct discrete states are, as long as only a finite number of states can exist. However, from the point of view of current human information managing activities, state granularity must be substantial in order to play a role in the economy of information handling. Discretization procedures designed to lead to i-flux saving processes have thus been ubiquitous in the

realm of information management (Suteanu 2010b). A practically successful way of addressing discretization consists of segmenting the enormous richness of the natural world to produce a limited number of categories, and, in some sense, to further manipulate every entity as if it were identical to all the others in the same class. The resulting "categorization" was mentioned by Plato and famously discussed in detail by Aristotle (1995) in his Organon. Since then, the subject of classification has been approached in many types of context and from innumerable points of view.

For large information managing systems that handle a wide diversity of informational components, classification as a form of discretization is essential. Classification extensively saves information processing capacities. On one hand, it allows the application of identical procedures to a whole group of entities, avoiding the need for the elaboration and execution of innumerable specific operations. On the other hand, in every "class" or "bin" many different entities are collapsed into a single one and further treated as one element. Applied iteratively on successive hierarchic levels, it provides a radically simplified way of considering rich collections of elements and their relations, and enhancing the understanding of the studied systems (for a review of attempts at classifying questions in science, see Dillon 1984).

3. Challenges and implications of classification and hierarchization

While classification-based discretization seems to be a necessity in the case of information-managing systems such as DE, it also involves major difficulties. Pointing out the instability inherent in classification criteria, George Perec sees classification as a futile and unavoidable human need to organize the world according to "a single code", which "will never work" (Perec 1999 cited in Maciel 2006). Max Planck (1936, 11-17) shows that classifications that are absolute and suitable to every purpose cannot exist; they all unavoidably involve arbitrary presuppositions. There is no "definite principle", in any science, able to lead to a structure "evolving from its own nature", and practical considerations are complemented by questions of value when an informational edifice based on classification is erected. One may metaphorically represent this situation as a configuration in n-dimensional space. If the studied properties are all constrained to one single dimension (n=1), the "only" task of classification is to decide upon the proper boundaries of the classes (which may be objective to a certain degree, which also depends on the features to be classified), as well as on the goals and criteria of the classifier. However, if the dimension n is higher than 1 (i.e. considering more than one property), finding a criterion for classification means, first of all, deciding upon the "axis" (or criterion) to be applied to divide the entities into classes. The degree of arbitrariness involved in classification increases with higher values of n, and the classification task becomes particularly challenging when, as it usually happens in the real world, n is large, or possibly (for practical purposes) infinite. The latter may actually be the case. The richness of the real world implies that no point of view may be "exhaustive", one can rarely address the "full information" about the environment (Parikh 2010, 39-41).

Classification systems rely on two sets of assumptions regarding the applied pre-established principles (Beghtol 1986): they are expected (i) to lead to "meaningful

relationships" among the resulting elements as well as between the latter and the approached objects in the world, and (ii) to be "beneficial to the users". Both of these assumptions are challenged in the context of DE. The relationships in (i) are subject to change – as a consequence of developments occurring on different levels, from progress in distinct disciplines to wider paradigm shifts. Due to the multitude of information input sources and procedures as well as the large variety of users who should benefit from information integration, producing changes to classifications is a very delicate task. Depending on the design of information integration, changes to classification criteria applied to already embedded, integrated information may become highly time- and energy-intensive and may even require major transformations in system design. Assumption (ii) is even more fragile, given the current heterogeneity in user goals, background, methods, expectations, etc., as well as the ways in which such heterogeneity can change in the future. Not just the value, but also the validity of classifications depends, in fact, on the group of users (Dupré 2006).

There are major dangers related to changes in classification criteria. As soon as classification occurs, all the elements in a class are considered identical to each other (at least for some points of view) and the same procedures are then applied to each and any of them. However, if different criteria are applied for different classifications, the groups of elements considered to be identical change. Object comparison may thus easily slip into confusion, which is avoidable only if classification criteria are made explicit and are perfectly well understood by all the users of information (which would be difficult to guarantee). Such change and confusion may have deep implications especially when classification is expected to have explanatory power (Atran 1981).

The discrepancy between the richness of the real world and the relatively small number of categories in desirable classifications has been addressed by the apparently simple and useful concept of hierarchization. It has been widely applied, from taxonomy to conceptual clustering (Gennari et al. 1989). Nevertheless, hierarchical classification is neither simple, nor exempt of challenges. The classification problems are carried over from one level to another. Moreover, attempts at an improved classification are discouraged: on one hand, the multiple levels of classification are less transparent and potential inadequacies have lower chances of being identified; on the other hand, even if possible improvements can be found, their implementation in a given hierarchical structure may require inconceivably deep and complex changes to the whole system. The problems of classification are critical to science, not only due to the fact that the resulting structure is not uniquely defined, but mainly because once such a "tree" is in place it may determine or at least influence future developments for a very long time. As Dupré (2006) notices, classified objects are often only distinguished after classification. Moreover, its very presence tends to obscure other (possibly better) structures; the "tree" might thus make scientists not see the "forest" of potential alternatives. The resulting tree might not be easy to redesign and replace.

Moreover, relationships within and among sets and classification categories, within the analyzed system and between the system and its environment are as important as the objects themselves - or even more so. Relations "are the glue that holds things together" (Parikh, 38). Insightful detailed explorations concerning relations and transformations between states have already been made by Frege (1966, 66-90). While relations are often

considered in the process of classification, their nature makes them often overflow the boundaries established between classification categories, and even between their hierarchical levels. For this reason, it might be correct to assume that relations should not be treated with the same cookie cutters applied to the delimitation of object bins.

4. Implications for Digital Earth

If the DE system should offer seamless integration of fast-accumulating, increasingly diverse information and effective, targeted access to information according to needs, then flexibility and adaptability will be among the main objectives pursued for its design and development. Paradoxically, flexibility and adaptability are among the most challenged features due to the problems arising from classification processes. These are not problems due only to imperfect – and thus improvable – procedures applied to classification. Beyond any correctible aspects of the ways categorization is performed, there are – by principle – obstacles to change. The classification structure tends to preserve its shape and reject even the identification, not only the implementation of major transformations. Classification schemes work like entities protected against i-flux.

There is thus a severe contrast between the crucial need for classifications, which keep i-flux low and make fast and effective information handling possible, and the strong inertia that classifications tend to oppose to the equally essential need for flexibility. There might be no quick solution to this problem. Perhaps identifying the problem and carefully considering its roots would be a first step in the right direction. However, useful ways of addressing it may already be at hand. A possible direction may concern the management of information before even classification procedures start.

One way of affecting information and moulding it by applying classifications, without building rigidity into the edifice, is to keep classification as external to the items being classified. Classification trees would thus be supplementary constructs, projecting links to the multidimensional entities themselves, without affecting them directly. In fact, any form of data processing implicitly assumes or induces a form of classification and contributes to the erection of a hard to change construction. Therefore, the initial form of collected data – which we may call "primitive information" – should always be preserved as such, irrespective of the operations to be performed later on. Any such operations could use copies of the initial datasets, transformed according to needs. Of course, "primitive information" is a relative term, since data collection relies on sets of assumptions, goals, methods, which all put their fingerprint on the outcome. What one can do is keep information in a form that is closest to the data collection stage.

While certain mechanisms and technical solutions will have to be found, the resulting information structure for Digital Earth could consist of "primitive" datasets and multiple, changeable and evolving sets of network constructions, pointing to relations and representing classifications performed according to identified needs. Data are understood here in the most general form, including sets of signs and concepts, which increasingly represent our objects of study (Bunn 1981, 170). New perspectives and fresh relations can lead anytime to the creation of other networks of relations and classifications, without affecting the actual data.

References

Atran, Scott. 1981. Natural classification. Social Science Information 20:37-91.

Aristotle. 1995. Categories. In Barnes, Jonathan. The Complete Works of Aristotle, transl. J. L. Ackrill. Princeton: Princeton University Press: 3–24.

Beghtol, Clare. 1986. Semantic validity: concepts of warrant in bibliographic classification systems. Library Resources and Technical Services 30, 2:109-125.

Bunn, James H. 1981. The Dimensionality of Signs, Tools, and Models. Bloomington: Indiana University Press.

Craglia, Max, Kees de Bie, Davina Jackson, Martino Pesaresi, Gábor Remetey-Fülöpp, Changlin Wang, Alessandro Annoni, Ling Bian, Fred Campbell, Manfred Ehlers, John van Genderen, Michael Goodchild, Huadong Guo, Anthony Lewis, Richard Simpson, Andrew Skidmore, and Peter Woodgate. 2012. Digital Earth 2020: towards the vision for the next decade. International Journal of Digital Earth 5:1, 4-21.

Dillon, J.T. 1984. The classification of research questions. Review of Educational Research 54, 3:327-361.

Dupré, John. 2006. Scientific classification. Theory, Culture & Society 23:30-32.

Fredkin, Edward. 2004. Five big questions with pretty simple answers. IBM Journal of Research and Development, 48, 1, 31-44.

Frege, Gottlob. 1966. Funktion, Begriff, Bedeutung. Fuenf logische Studien. Goettingen: Vandenhoeck & Ruprecht.

Gennari, John H., Patrick W. Langley, Douglas H. Fisher. 1989. Models of incremental concept formation. Artificial Intelligence 40: 11–61

Gore, Albert. 1999. The Digital Earth: understanding our planet in the 21st century. Photogrammetric Engineering and Remote Sensing 65, 5: 528.

Maciel, Maria Esther. 2006. The unclassifiable. Theory, Culture & Society 23:47-50.

Parikh, Prashant. 2010. Language and equilibrium. Cambridge, USA: MIT Press.

Parsons, Mark A., Oystein Godoy, Ellsworth LeDrew, Taco F. de Bruin, Bruno Danis, Scott Tomlinson, David Carlson. 2011. A conceptual framework for managing very diverse data for complex, interdisciplinary science. Journal of Information Science 37, 6:555-569.

Perec, George. 1999. Species of Spaces and Other Pieces. Trans. John Sturrock. London: Penguin Books.

Planck, Max. 1936. The philosophy of physics. New York: W.W.Norton & Company.

Suteanu, Cristian. 2007. You cannot swim in foam: Information flooding and information dilution in processes of change. International Journal of the Humanities 5: 69-75.

Suteanu, Cristian. 2010a. A scale-space information flux approach to natural irregular patterns: methods and applications. Journal of Environmental Informatics. 16, 2: 57-69.

Suteanu, Cristian. 2010b. Towards an environmental science centred on informational processes: artefacts, information dynamics, and the analogue-to-digital transition. In Explorations in the Philosophy of Engineering and Artefact, ed. Viorel Guliciuc. Cambridge: Cambridge Scholars Publishing, 167-178.

Suteanu, Cristian. 2011. On information and the geographical integration of information processes. Biocosmology–Neo-Aristotelism 1, 2/3: 167-180.