HAL
open science

Big Data: How Geo-information Helped Shape the Future of Data Engineering

Robert Jeansoulin

To cite this version:

Robert Jeansoulin. Big Data: How Geo-information Helped Shape the Future of Data Engineering. Barry Wellar. AutoCarto Six Retrospective, Information Research Board Inc., pp.190-201, 2013, 978-0-9921435-0-3. hal-00950806

HAL Id: hal-00950806

https://hal.science/hal-00950806v1

Submitted on 22 Feb 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

(published in: AutoCarto Six Retrospective, Barry Wellar, pp.190-201)

Big data: how geo-information helped shape the future of data engineering

Robert Jeansoulin

CNRS senior researcher, Université Paris-Est Marne-la-Vallée

Laboratoire d'informatique, Institut Gaspard Monge

(on leave, as: Science Advisor at French General Consulate in Quebec)

robert.jeansoulin@univ-mlv.fr

ABSTRACT. Very large data sets are the common rule in automated mapping, GIS, remote sensing and what we can name geo-information. Indeed, in 1983 Landsat was delivering already Giga-bytes of data, and other sensors were in orbit or ready for launch, and a tantamount of cartographic data was being digitized. The retrospective paper re-visits different issues that geo-information sciences had to face from the early stages on. First challenge, Structure: to bring some structure to the data registered from a sampled signal (metadata). Processing: so huge amounts of data, required big computers and fast algorithms. Uncertainty: describing the kind of errors, and their quantification. Consistency: when merging different sources of data, it is mandatory to question if we were allowed to merge them. Ontologies: the comparison of geo data, over space and over several decades now, imposes clear and shared definitions, if any kind of decision should be built upon them. All these issues are the background of Internet queries, and the underlying technology has been shaped during those years were geo-information engineering emerged.

KEYWORDS. Automated mapping, remote sensing, GIS, Big data, machine learning, data quality, geo-information, knowledge systems, ontologies, exploratory data analysis.

A. Title of AutoCarto Six paper. Automatic cartography of agricultural zones by means of multi-temporal segmentation of remote sensing images.

B. Reason for paper?

In 1983, fresh doctor working on the above title, the purpose of my then sponsor[1] was to prepare the launch of satellite SPOT and the forthcoming commercialization of its products, with a focus on vegetation monitoring.

It may look unbelievable today, but we were not equipped with image-capable screens, only alphanumeric consoles: everything had to be printed for being displayed. However, data were there, big matrices of data that couldn't be turned easily into images.

[1] Centre National d'Etudes Spatiales, Département de Traitement de l'image, Toulouse, France

Therefore, we were forced to crunch data, failing to be able to look at them! The amount of images that satellites such as Landsat were able to harvest, was phenomenal. Giga, Tera bytes of pixels: we were using the super computers of that time.

Dealing with "big data" before the term was popularized? The armory of mathematical tools also was almost there: principal component analysis, multi-dimensional correlation, template matching, and so on. We may say that in remote sensing, in photogrammetric engineering, in geographical information systems, or for short in geo-information (or geomatics), we have pioneered what is named today *Big data*.

This paper browses the principal issues that the geo-information science had to face from the early stages on. First challenge: to bring some structure to the data registered from a sampled signal, what eventually gave metadata and the ability to merge images into large retrieval systems in the Internet. Also: processing for so huge amounts of data, required big computers and fast algorithms. Data uncertainty: the description of the kind of errors, and their quantification was necessary from the very beginning. Data consistency: as soon as we started merging different sources of data, it became mandatory to question if and how far we were allowed to merge them (in French we say: *mariage de la carpe et du lapin*, carp and rabbit wedding).

Finally, "ontology questions" are addressed here, because the comparison of geo data, piled up for several decades now, and all around the globe, imposes clearer definition on what they are data of, what comparison can be made, what real evolution or differences they measure, and what kind of decision can we built upon them.

Today, these same longstanding issues and questions continue to be raised in the context of big data.

C. Data structure: From ancillary data to Metadata

Remote sensing imagery inaugurated the use of metadata as soon as the first images had to be registered, overlaid, and when different sensors were used.

Basically we have only two kinds of data: words and numbers. And if we have a number, we do need words too: if 24, 24 what? Natural or physical sciences deal with numbers, with signal or image processing, with geometry, time series, etc. But text processing is the main approach in Big data: "*In principio erat Verbum*", the very first words of the Gospel of John. The two didn't fit well 30 years ago, but nowadays the difference is not so drastic.

Noticeably today is the widespread use of metadata. I do remember that we didn't use the word metadata but "ancillary data", to name the data associated with remote sensing imagery: for instance the ground resolution, wavelengths, etc. (see below). The name change denotes an upgrade for something secondary (*ancillary*) to a more important role (*meta*). The way we consider data has evolved as well.

For instance: pixel data. A pixel is the approximation of a ground surface, from which we measure a reflected (or emitted) radiometry within a certain wavelength range, integrating diffraction effect, absorption, etc. As many of us, I spent a lot of time modeling these effects, filtering them to improve the data from a status of "raw data" to a status of "corrected data". Also, pixels aren't processed one by one, but as a statistical variable that receives class membership or geometrical properties (e.g. border pixel). These properties must be described and registered into some structure: a "processed image" has a lot of such attached information (metadata).

Libraries were confronted to the metadata issue and developed *MARC* in the sixties, a markup language to which HTML owes a lot. Next step, the *Dublin Core* in 1995. In automated cartography, one big question was: how to introduce the topology in the data vector representation: it's implicit from the geometry, but the burden of re-computing it is much too heavy. Then, in the 90's several NGOs[2] were working on what became the *ISO 19101: 2002 Geographic information Reference model*.

For instance, this table represents the geometry and the topology of a set of land parcels and allows determining that the union of 2 and 3 forms a single hole into 1.

#	Coordinates (or vertices)	contains	is in	touches	has hole	… more …
1	*x,y; x,y; x,y ; x,y; x,y …*	**2;3**	**-**	**-**	1	…
2	*x,y; x,y; x,y; x,y …*	**-**	**1**	**3**	0	…
3	*x,y; x,y; x,y …*	**-**	**1**	**2**	0	…

The content of such tables is described in the ISO reference model: the polygon-arc-node model, with all topology relationships. Moreover, semantics and rule can be added too (see: data consistency).

In Big data, there are trillions of sparse, scattered, unrelated docs (unstructured) anywhere on the Internet, and the goal of so called "robots" is to attach them some structure (the indexing process) and trying to rank them in response to a single request (querying the Internet) or to a complex and multi-morphed request (data analytics).

The concept of unstructured data wasn't in use 30 years ago. Relational databases were just on the rise (and still are embedded in most data engineering). The term *semi-structured data* appeared in 1995 in the community of database specialists, and XML was first established in 1997 (built upon the SGML of the late 80's).

Hence, there are reasons to consider that data processing of signal and image is one of the many precursors of today big data.

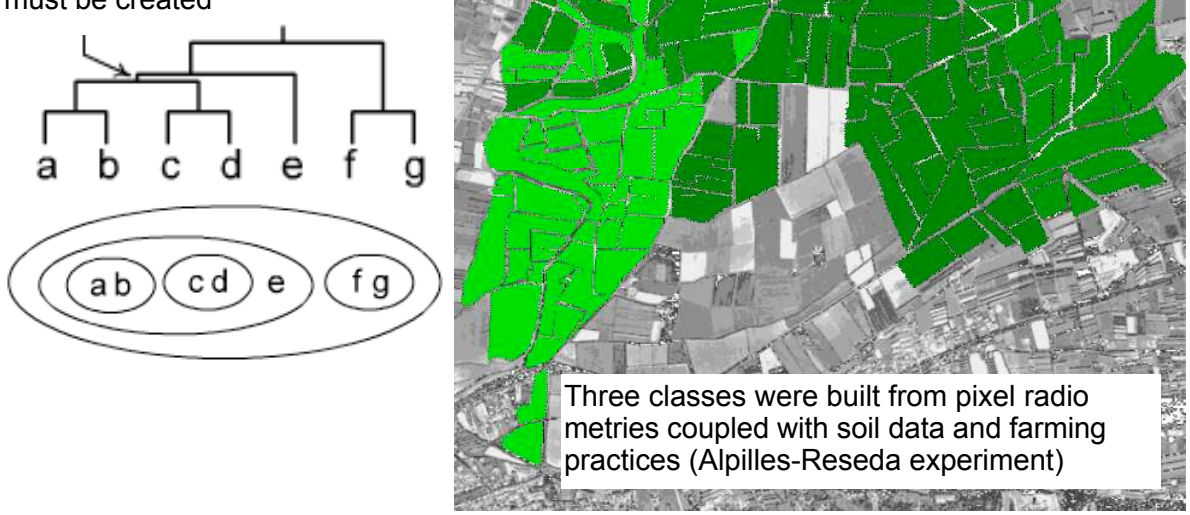
D. <u>Data Processing</u>: From data analysis to data mining

Signal processing has been a field of experimentation of innovative statistical, stochastic or data transformation approaches, as well as intensive computational methods and efficient algorithms for the processing of very large data sets. Exploratory data analysis (EDA) is linked with Tuckey and Fourier analysis and similar approaches,

[2] National geographic offices, such as Ordnance Survey, IGN Geomatics Canada, USGS, etc.

developed mainly in the field of signal processing. The term is outdated now, but it certainly contributed to foster the research in high-end computing. Let's recall for instance, the hierarchical classification algorithms used for pixel classification in Landsat images [Jeansoulin 1981], and the use of more sophisticated approaches for unsupervised classification, such as the "dynamic clusters" [Diday 1973].

Figure 1. Hierarchical clustering example (sometimes referred to as phylogenetic tree).



The descending operation tries to split the data domain into relevant sub-classes at each level, leading to a tree structure (see section: Ontologies). It is then applied to field parcels according to their median or average pixel value [Olioso 1998].

The expression *data mining* traces back to the 90's, familiar to database scientists and mainly as an operational approach of machine learning whose foundations are linked to the Turing machine: the field was more theoretical, between logics and computing languages, lambda-calculus. Support vector machine (SVM) algorithms were developed as non-probabilistic binary linear classifier [Vapnik 1995]

Nowadays, these terms have been more or less wrapped up in the successive buzzwords of business intelligence, data analytics, and big data. In spite of other differences, the fact is that the mainstream shifted from signal processing to the world of Internet and e-commerce. But, look deep into the algorithms: the geo-processing legacy is there. From a computational point of view (supercomputer and huge data storage), or with respect to the underlying exploratory approaches, geo-processing has contributed to pave the way.

E. <u>Data uncertainty</u>: From precision to quality indicators

Geo-information deals with the “real world”: without arguing about philosophical truth, it deals at least with a same and single “world” which can be modeled and measured by different means, at different scales, from multiple viewpoints … and everything has to be consistent (*logically* consistent). Therefore, geo-processing talks not only about a particular aspect of reality, but also about a consistent framework for this reality.

At first, measurements quality was merely limited to precision: what particular location on Earth a particular pixel represents? What is the real radiometric contribution of this location actually represented in the pixel value? Image registration, sensor fusion were the first and most difficult tasks at the time. Soon after that, the confidence on data classification was the big issue. How much wheat the USSR is really harvesting? Such kind of questions were investigated by the relative same amount of computing power than what the NSA can gather today for processing trillions of emails.

When it became easier to merge remote sensing imagery and geographical databases, much more complex questions were at hand. The gain in ground resolution enabled us to reach new terrains: from the large rectangular crops of Middle West to small vineyards in Provence, and now to urban roof gardens. Sociology is no longer some weird science taught to hippies, but a science we can have trans-disciplinary talks with, as we did with Agronomy.

The challenge about quality is no longer about data precision (though still is), but about consistency: are we talking about the same thing when merging different data.

Geo-processing has extensively investigated the question of quality, and a bunch of quality indicators has been designed and the quality domain has been structured by international consensus:

Since the 90’s, geo-information specialists met in consortia such as OGC, and eventually established an ISO technical Committee (ISO TC211) to discuss and publish standards for geo-information (geomatics): data models and data quality information were among the most specific outcomes: ISO 19101: “reference model”, and ISO 19113: “quality principles” were finally issued in 2002, reflecting a common ground between the various national geographic organizations. The national statistics agencies were closely working with these specialists, because the same quality issues are important for the countries and for international comparison as well. International bodies, such as Unesco, OECD, Eurostat, are also aware of the same issue since many years, but the automation of cartography was probably among the very first pioneers.

The next table represents the actual consensus about metadata for quality in ISO2002.

Data quality element / sub-element	Description
Completeness (omission,	Presence of features, their attributes and relationships

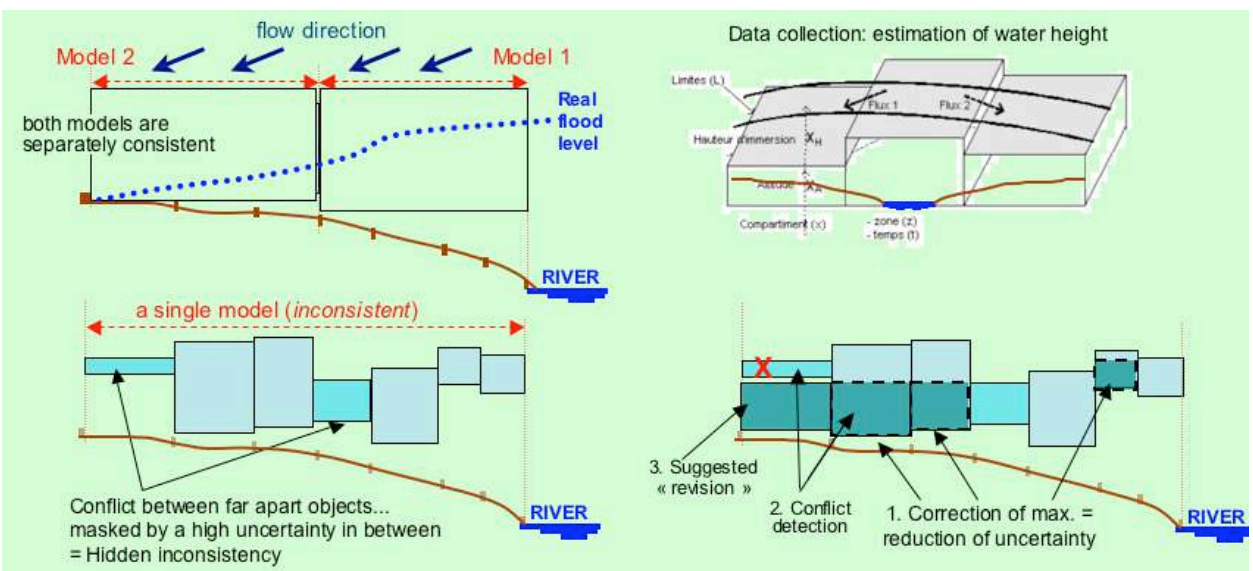
commission, logical consistency)	(absence or excess of data, adherence to rules of the data structure)
Conceptual consistency	Adherence to rules of the conceptual schema, to value domains, etc.
Topological consistency	Correctness of explicit topology, closeness to respective position of features
Positional accuracy	Accuracy in absolute point positioning, gridded data positioning
Temporal consistency	Accuracy of temporal attributes and their relationships
Thematic accuracy	Accuracy of quantitative attributes, class correctness

F. Data consistency: uncertain but rational knowledge

Understanding the many causes of data uncertainty sheds light on the many approximations made all along the process of gathering and measuring even the simplest datum: for instance a ground temperature.

Considering that data are always somewhat inexact, and considering that data are always depending on a model imperfectly representing a reduced aspect of the reality, it is important to provide guidelines or constraints. Every time we can provide some constraints, we can confront the data, and issue a warning for each detected conflict.

Here below is a snapshot to a successful experiment developed during the FP5 European project REVIG!S[3]: how to revise very uncertain flood data using direction of water flow as constraints [JeansoulinWilson 2002].

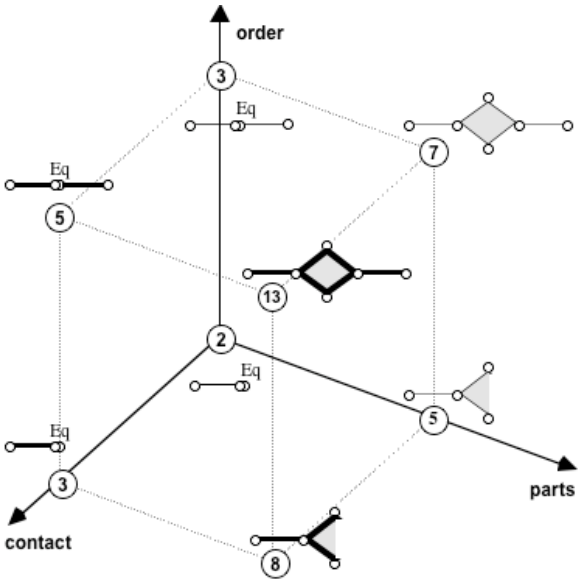


In Big data, the "prediction" of the annual flu wave by looking through Internet "medical queries" has got a lot of media attention. It isn't necessarily a convincing example, but the archetypical "analytics" story is IBM Watson, when it overcomes two Jeopardy champions, in 2011. The DeepQA project behind Watson is making intense use of geo-information reasoning for answering questions such as "*They're the two states you could be reentering if you're crossing Florida's northern border*" [Ferrucci 2010].

[3] REVIG!S involved universities of Marseilles, Leicester, Keele, Technical Vienna, Pisa, Twente (ITC), Laval

Mereotopology, region connectedness calculus (RCC) or Allen's interval algebra [Allen 1983], developed in the 80's, are extensively used today for constraining queries and making answers more narrow and efficient.

The next figure gives the lattice the 8 possible such algebras [Euzenat 1997].



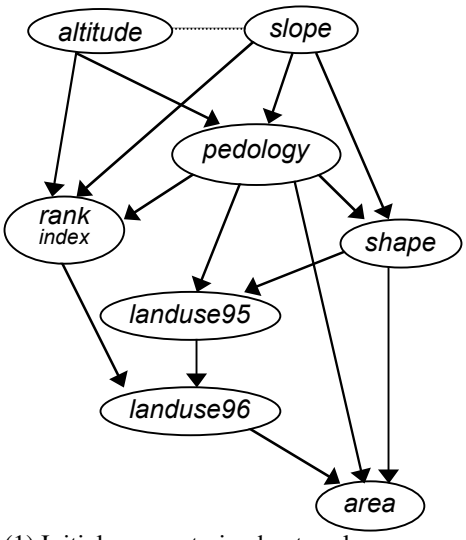
Lattice of 8 mereotopology algebras: each node gives the number of allowed relations per algebra (from 2 to 13); and the axis have the following effects:

contact introduces <u>arc</u> (bold) relations,

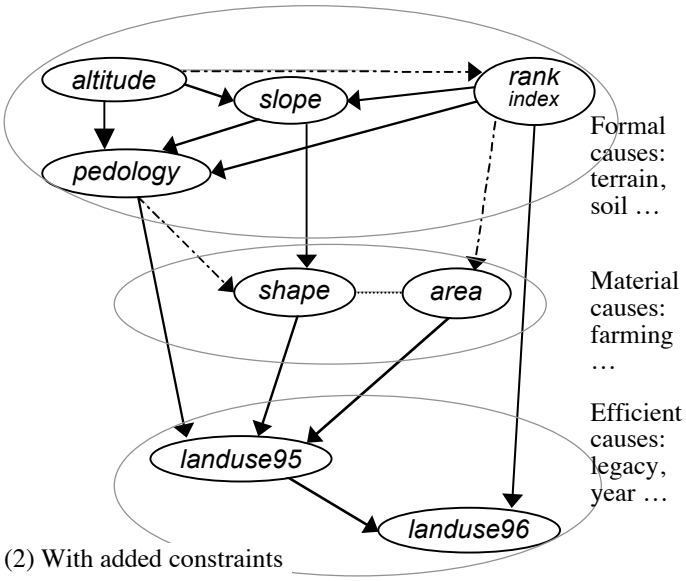
parts explodes the <u>equivalence</u> relation Eq,

order unwraps the graph around a <u>symmetry axis</u> (Allen).

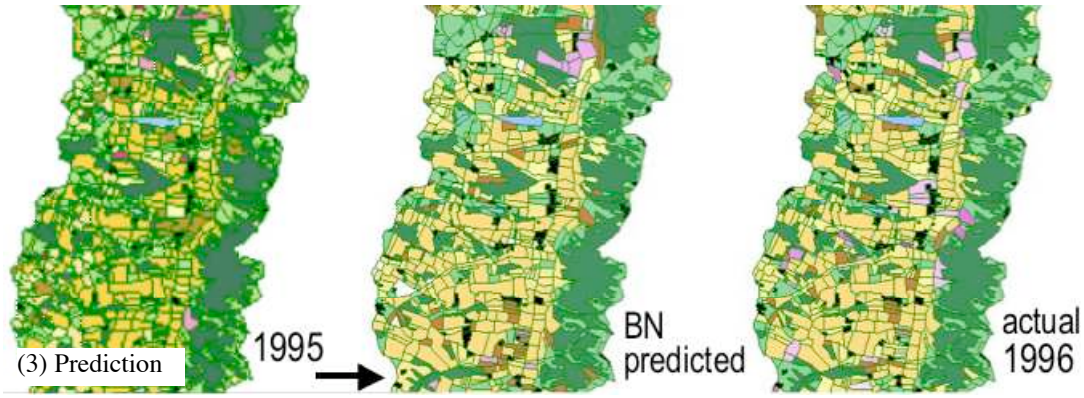
Reasoning under constraints marries well with stochastic reasoning, as with Bayes networks [Cavarroc 2004]. The constrained network (2) is used to build a prediction for the following year, from the land use in 1995.



(1) Initial, unconstrained network



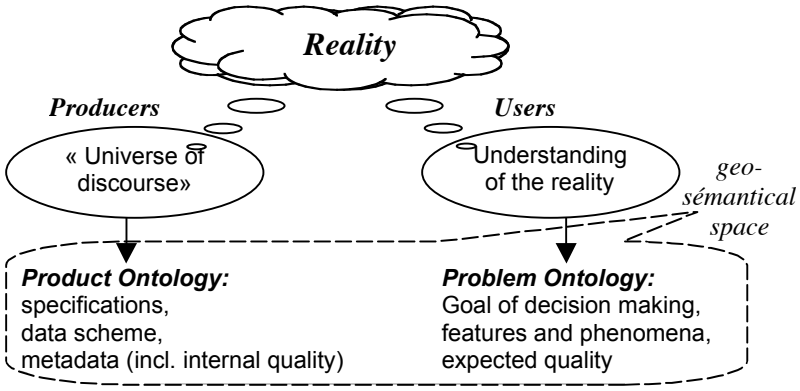
(2) With added constraints



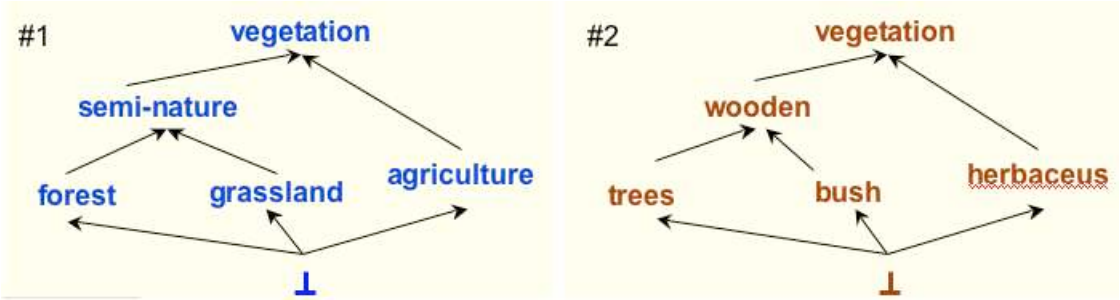
G. <u>Ontologies</u>: Data are Acts, not Facts

Geo scientist are still classifying and zoning, but much more attention is turned to the meaning of the process, the interpretability of the result, and the ability to use it within a decision process. Moreover, geo-information also raised the question of what is in data, common agreement, external quality, data usability, etc. different aspects of a more general question often summarized into the word “ontologies” (plural!), and the subsequent problem of “ontology alignment” [Gruber 1994] [Halevy 2005].

Here again, the terminology has been brought to public attention by the widespread of the Internet and the mainstream studies following it, but here again, geo-information was developing its research when questioning the problems of global data quality, of managing different quality levels for different types of information, and geomatics was among the first to provide a systematic and comprehensive approach, with the above-mentioned ISO 191xx series of standards, in particular in 2002 with ISO 19150: “ontologies”, and ISO 19157: “data quality”. The book [Devillers 2006] regroups a good part of this research on ontologies and quality, in particular the notion of external quality (quality for the user) as illustrated here.

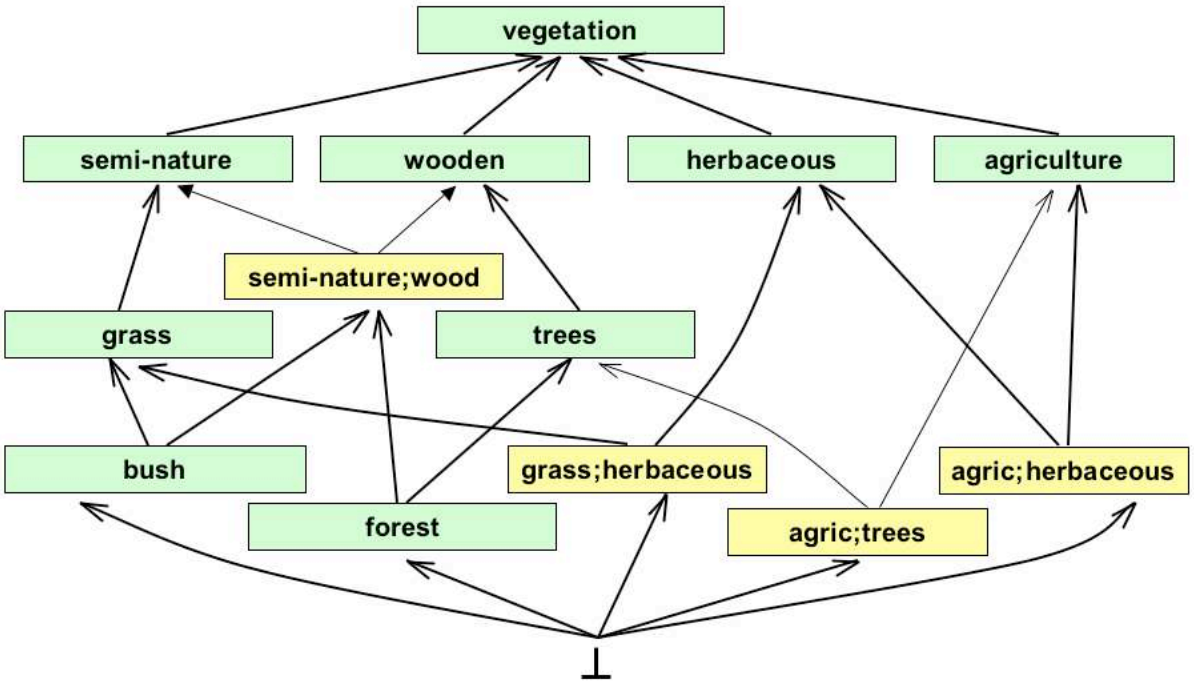


Let’s develop an example of ontologies alignment with agricultural data. Given 2 graphs representing two different surveys of the same region [Pham 2004]:



when the same locations (parcels) receive values from both ontologies, we can build a Galois lattice summing up the fusion of information. Then, we notice that some nodes (yellow) must be defined as a consensus between the two original ontologies,

such as “grass;herbaceous”. It leads to a compromise between the two ontologies, but much more efficient and meaningful than the mere cross product of the ontologies.



Therefore, ontologies are not made solely for the Internet, but for questioning “what there is”, as W. O. Quine was saying about the Ontology (the philosophy Ontology).

I. <u>Conclusion</u>

Some 30 years ago, the conjunction of the multiplication of remote sensing imagery (Landsat 1976, then SPOT), the early stages of desktop image processing, the automation of the cartographic processing, and a rapidly increasing computing power

(the Moore law, version 2 was published in 1975[4]), offered an opportunity to collect, merge and process together an enormous amount of data, more huge than ever collected and processed by machines. The engineers and researchers involved in geo-information were consequently on the leading edge for the development of tools and methods that eventually are part of what today is termed: big data.

Finally, in closing, a tribute.

It is appropriate to recognize the expertise that geo-information specialists have developed in data engineering. Decision-makers are increasingly relying on numbers to prepare and make their decisions. However, it appears that few of them are aware of the way these numbers are produced, that is, data are the result of many small decision-making processes at all stages. Data are not Facts, but Acts.

From the signal to the semantics of information, the science of geo-information has confronted, and been confronted by, a very large range of issues, and has brought its contribution to many data models, and many algorithms.

Dealing on an everyday basis with data and models, but dealing basically with the real world, the geo-information scientist must continue to seek out the right tools and representations, and thereby continue to pioneer advances in data engineering.

J. References.

A. Halevy. 2005. Why your data don't mix?. *ACM Queue*.

Gruber, T and Olsen, G, 1994, An ontology for engineering mathematics. *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning,* 258–269.

E. Diday. 1973. The dynamic clusters method in nonhierarchical clustering. *International Journal of Computer & Information Sciences*, March 1973, Volume 2, Issue 1, pp 61-88

Corinna Cortes, Vladimir N. Vapnik, 1995. Support-Vector Networks. *Machine Learning*, 20, 1995. http://www.springerlink.com/content/k238jx04hm87j80g/

Devillers, Rodolphe; Jeansoulin, Robert. 2006. *Fundamentals of Spatial Data Quality*. 2006. ISTE Publishing Company, London UK.

Jeansoulin, Robert; Papini, Odile; Prade, Henri; Schockaert, Steven. 2010. *Methods for handling imperfect spatial information*. 256pages. 2010,Springer Berlin Heidelberg.

Degenne, Pascal; Lo Seen, D; Parigot, Didier; Forax, Rémi; Tran, Annelise; Ait Lahcen,

[4] Computing power is "*doubling every two years*": in *Progress in digital integrated electronics*, Moore, G.E., 1975, IEEE.

A; Curé, Olivier; Jeansoulin, Robert. 2009. Design of a domain specific language for modelling processes. Landscapes. *Ecological Modelling*, 220(24), 3527-3535, Elsevier

Edwards, Geoffrey, Jeansoulin, Robert. 2004. Data fusion: from a logic perspective with a view to implementation. *International Journal of Geographical Information Science*, 18(4),303-307. Taylor & Francis.

Jeansoulin, Robert; Wilson, Nic. 2002. Quality of geographic information: Ontological approach and artificial intelligence tools in the Revigis project. *EC-GI& GIS Workshop*. 12,2002,

Jeansoulin, Robert; Fontaine, Yves; Frei, 1981. Werner. Multitemporal segmentation by means of fuzzy sets. *7th LARS Symposium on Machine processing of remotely sensed data, with special emphasis on range, forest, and wetlands assessment*. 336-340,1981,Purdue University.

David Ferrucci, et al. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, Fall 2010. publisshed by AAAI

Michael Goodchild, 2010. Twenty years of progress: GIScience in 2010. *Journal of spatial Information Science* JOSIS (1), pp. 3–20

James F. Allen: *Maintaining knowledge about temporal intervals*. In: *Communications of the ACM*. 26 November 1983. ACM Press. pp. 832–843.

Marie-Aline Cavarroc, Salem Benferhat, Robert Jeansoulin, 2004. Modeling Landuse changes using Bayesian Networks. *22nd IASTED Intl. Conf. Artificial Intelligence and Applications*, Innsbruck, Austria.

Albert Olioso et Al. 1998. Spatial aspects in the Alpilles-ReSeDA project. *Scaling and Modeling in Forestry: Application in Remote Sensing and GIS*, D Marceau (Ed.), Université de Montréal, Québec pp 92–102.

Jérôme Euzenat, Christian Bessière, Robert Jeansoulin, Joel Revault, Sylviane Schwer, 1997. Raisonnement spatial et temporel. *Bulletin de l'Association Française pour l'Intelligence Artificielle, vol.29,* 2-13.

TT Pham, V Phan-Luong, Robert Jeansoulin, 2004. Data Quality Based Fusion: Application to the Land Cover. *7th International Conference on Information Fusion* (FUSION'04)