Reprinted from

Seventh International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Range, Forest and Wetlands Assessment

June 23 - 26, 1981

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

PROBLEMS ON DATA STRUCTURATION ABOUT PALEONTOLOGICAL COLLECTIONS

DANIEL PAJAUD, MARIE-JOSE ROULET

Université Pierre et Marie Curie
Paris, France

I. ABSTRACT

The paleontological collections of the University Pierre et Marie Curie (more than 800 000 samples, 4×10^6 data) present a complex structure relevant to nine themes : Classification , Stratigraphy , Geography , Environment, Paleontological Material, Collecting, Bibliography, Storage, Exploitation. The complexity of the data requires a sophisticated model of data structure for archiving in a data bank.

II. ABOUT PALEONTOLOGICAL DATA

A. BIOLOGICAL, GEOLOGICAL, FORTUITOUS DATA

Data are all the pieces of information, either qualitative and quantitative, objective or not, which are tied to each specimen.

- Some are linked to the living world and to its evolution through the fossiliferous ages during more than half a billion years. For example, the morphological characteristics of species, their phyletic relations. These data are expressed with words of the Terminology (descriptive terms), of the Nomenclature (names of living organisms), of the Taxonomy (names showing the place of organisms within the classification).

- Other data are a result of the geological history of our planet. For example, the paleogeographical characteristics of a previously marine province, the sedimentological characteristics of a deposit area, the characteristics linked to burying and fossilisation, the age of the referred to organisms and of preserving sediments. The words expressing these data are terms of Geography (localities, co-ordinates), terms of Sedimentary Petrography (names of rocks and formations), terms of Taphonomy (description of the state of fossil preservation) and terms of Stratigraphy (names of the geological stages).

- Lastly, some data are simply fortuitous and are a consequence of the paleontologist intervention. For example, the number of specimens of one sample, its place in the classification, its duplication or the scientific publications.

Potentially, about fifty items of data can be considered for each specimen.

B. CONCRETE AND ABSTRACT DATA

How do the data arise ?

- Concrete data example for a sea urchin :
  + preserved part (theca)
  + state of fragmentation (fragment)
  + state of resistance (brittle)
  + diagenesis (silicified)
  + other data (sedimentary information)

- Abstract data example for a frog (complementary characteristics which are linked with the scientific works and classifications) :
  + place of collection (U.P.M.C., Paris)
  + species (*Triadobatrachus massinoti*)
    ancient name (*Protobatrachus massinoti*)
  + author and date of creation (Piveteau , 1937)
  + stratigraphical level (Lower Trias)
  + place of discovery (North of Madagascar)
  + reference to published illustration (Piveteau, 1951, p.58, fig.35)
  + nomenclatural status (holotype)
  + collection number (S 2833 6B)

Part of this basic data is usually found on documents called "storage cards" to be stored in a card index. Such cards generally list other data including hierarchized abstract data, so called as it depends on certain basic data of hierarchized type.
Here is an example, in taxology, from a species of sea urchin, in the ascending order (from most precise to most general) :

+ species	*Micraster fortini*
+ genus	*Micraster*
+ family	*Micrasteridae*
+ order	*Spatangoidea*
+ class	*Echinoida*

But the information here concerns only a part of potential and actual information : this information is therefore incomplete.

Ideally, all the aspects mentionned have to be taken into account for the realization of the data structure. Each data has a long and complex history, it has been distorded by partial loss of information, appearence of doubts between two choices of indications which are mutually exclusive, appearence of fuzziness through new interpretations.

C. RELATIONSHIP AMONG PALEONTOLOGICAL DATA

Relationship exist among all these data, but they are not all of the same nature and they do not present the same degree of immediate interest.

- Some are of hierarchical type, for example all the data concerning Taxonomy (a species belongs to a genus which belongs to a family which belongs to an order...).

- Other are contingent and result from the logics of the history of the planet. For example, the relationship between a fossil, the strata which preserved it and the place where it was gathered : relationship which show some natural coincidence (such a species lived at such a time in such a place).

- Lastly, other relations are hazardous. For example, the haphazard of a discovery during a mission, the premises where the fossils are placed, the documentation state...

III. ABOUT DATA STRUCTURATION

A. H.B.D.S. SYSTEM

The "Hypergraph-Based Data Structure" (HBDS) is a data structure model based upon graph theory, on set theory and hypergraph concept.[1] This model appears as a universal tool for data structuring and storing, more especially in complex scientific applications as paleontological data processing.[2,3] This application presents each of hierarchical, network and relational aspects at the same time.
In HBDS System, the characteristics linked to the specimens are distributed among four abstract types of data : object, class, attribute, link.[1] In paleontological application, these data are distributed among nine thematical sets, nine hyperclasses.[4]

B. HYPERCLASS "CLASSIFICATION" (fig.1)

Taxonomy concerns the Life Sciences and constitutes the first thematical set. It deals with names given to different categories of living beings (fossils or living things) within a hierarchized classification system (taxology), which itself lies beside other systems in relation to behaviour, evolution and the life environment... forming an even larger category (the systematic). As for the storage of samples, this is specific to any system of physical storage of samples, but

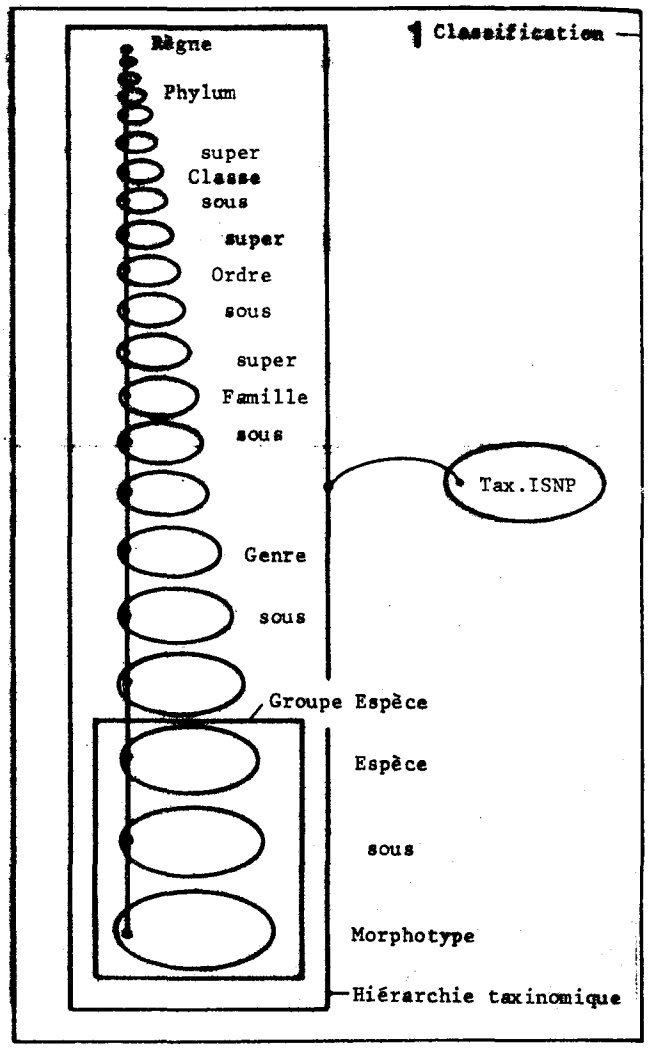


Fig.1 The hyperclass "CLASSIFICATION" A hierarchical substructure.

with certain pecularities proper to fossils. This first hyperclass includes 22 classes, whose 21 constitute a hierarchical substructure, with the class "KINGDOM" at the top and the class "MORPHOTYPE" at the base. Each class is the predecessor of a smaller class (except at the base) and is the successor of a larger class (except at the top).
Within this hierarchical hyperclass, three classes ("SPECIES", "SUBSPECIES" and "MORPHOTYPE") constitute a smaller hyperclass because special properties : relationship with "STRATIGRAPHY" and other thematical sets.

C. HYPERCLASS "STRATIGRAPHY" (fig.2)

In the second thematical set, we have five smaller sets (hyperclasses "CHRONOSTRATIGRAPHY", "BIOSTRATIGRAPHY", "LITHOSTRATI-

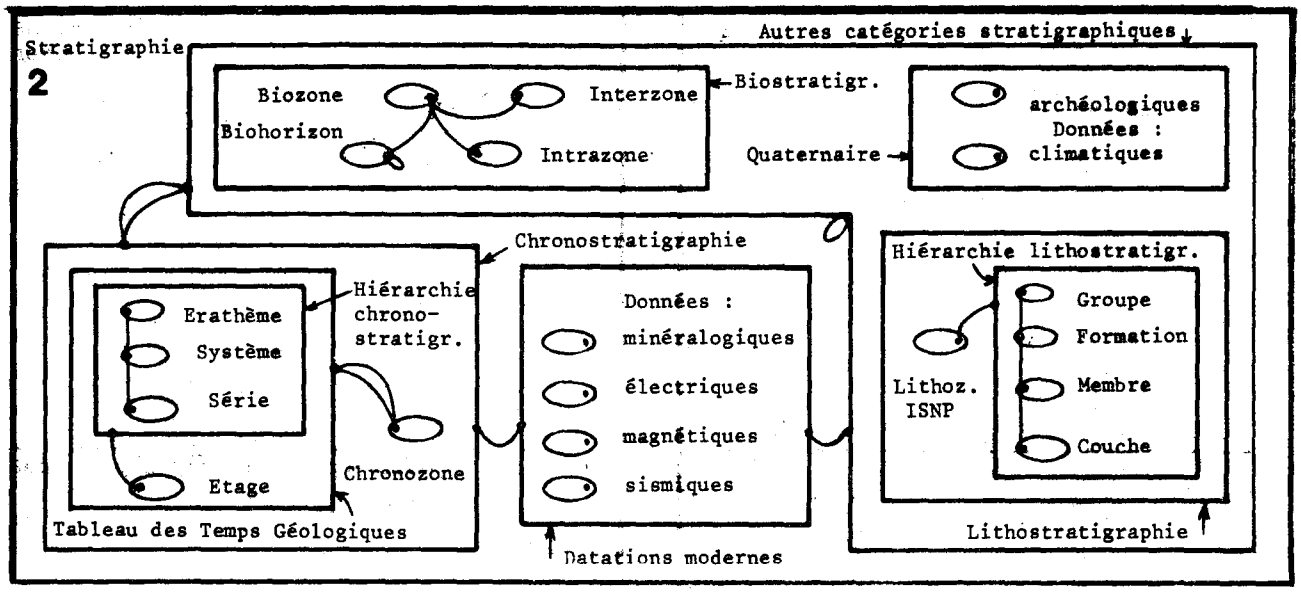


Fig.2 The hyperclass "*STRATIGRAPHY*". A hierarchical substructure with degenerated arborescence (isolated class) and nested hyperclasses, and non - hierarchical substructures with isolated classes.

GRAPHY", "*QUATERNARY*", "*MODERN DATATIONS*"), including 20 classes. This example brings a new element in relation to the previous one in the presence of a degenerated arborescence (which appears under the form of one isolated class : "*STAGE*") beside the functional arborescence (classes in cascade).
Another characteristic is the presence of relational but non hierarchical substructures. In fact, the classes are never isolated entirely. The objects can indeed justify in an effective way a certain number of relationship borne by links between classes. So we shall consider the multigraphs of links between classes within the hyperclass "*CHRONO-STRATIGRAPHY*". They are 22 links between 5 classes and the cobwet that represents them cannot be grasped or easily manipulated, neither by the curators nor by the students. The hypercomponents have then their entire justification : only three hyperlinks (with two supplementary hyperclasses). Within the complete hyperclass "*STRATIGRAPHY*", 9 hyperlinks replace a few hundred links. The degree of average conciseness of the hyperlinks in such an application is superior to 20.
Naturally, they are numerous links between "*STRATIGRAPHY*" and "*CLASSIFICATION*", but only 2 hyperlinks.

D. HYPERCLASS "*GEOGRAPHY*" (fig.3)

The third thematical set is the hyperclass "*GEOGRAPHY*", which includes 9 classes. One hyperclass is a hierarchical substructure (7 classes with "*CONTINENT*" at the top and "*NAMED PLACE*" at the base). But a new characteristic is the presence of intersected hyperclasses (in addition to nested substructures).

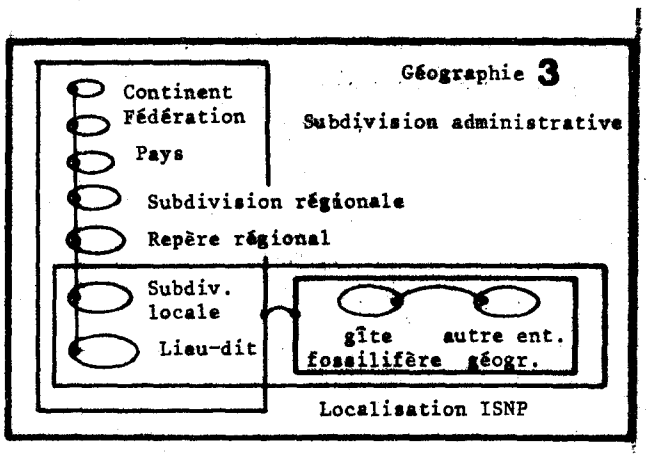
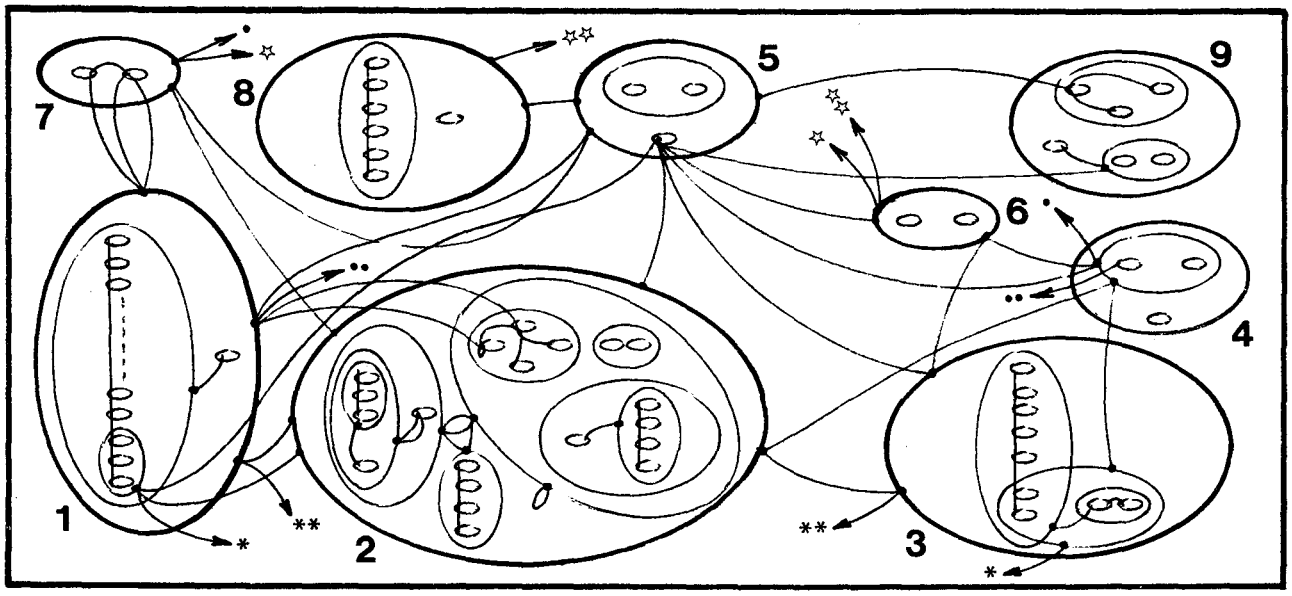


Fig.3 The hyperclass "*GEOGRAPHY*". Substructure with nested and intersected hyperclasses.

E. COMPLETE STRUCTURE (fig.4)

The skeleton structure (at the present time of its elaboration) is constituted by 74 classes and 13 hyperclasses, 8 links and 43 hyperlinks (which replace more than one thousand links). On 1977, only 47 classes and 15 hyperclasses, 3 links and 32 hyperlinks (for 650 links).[4.3]

It is important to note that as there are more than one thousand potential relations borne



1. "CLASSIFICATION" 2. "STRATIGRAPHIE" 3. "GEOGRAPHIE" 4. "ENVIRONNEMENT" 5. "MATERIEL PALEONTOLOGIQUE" 6. "RECOLTES" 7. "BIBLIOGRAPHIE" 8. "RANGEMENT" 9. "EXPLOITATION"

Fig.4 Skeleton structure of the paleontological application (nine thematical sets)

by links between classes, a given object meets only very few of these relationship.
In addition to links, we must consider attributes and the extension in their treatment. The attributes have led us to the part that can be qualified as "relational" in the HBDS model according to the meaning of "relational model". Nevertheless it is obvious that the links have a much more relational character than the attributes : an attribute corresponds to a characteristic which is a big restriction to the concept of relationship.
An extension in the treatment of the single attributes allows the notion of compound attributes to appear and requires the use of hyperattributes, which are themselves single or compound. For instance in the set *"PALEONTOLOGICAL MATERIAL"* (fig.5) :

+ single attribute : number of specimens
+ compound attribute : state of preservation
+ compound hyperattribute : availability

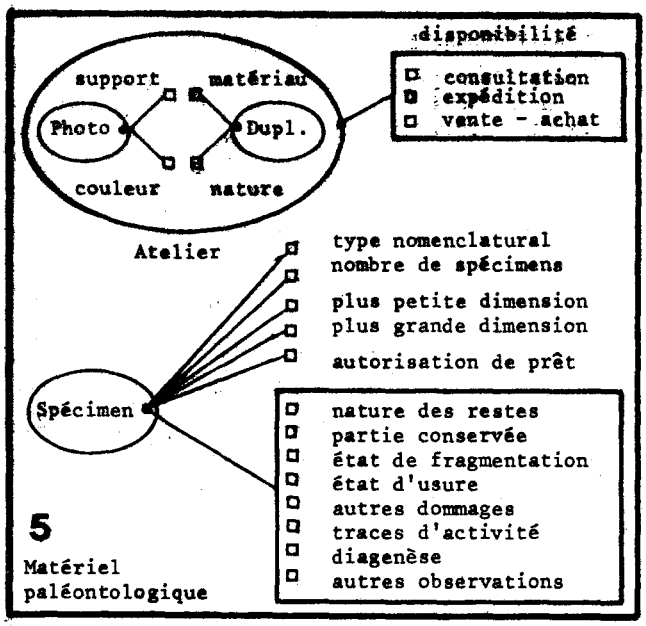


Fig.5 (Hyper-)classes attributes

Iv. IMPORTANCE OF FUZZINESS

A. DIFFERENT CASES OF FUZZINESS

Numerous cases of combined fuzziness have been discovered and listed.[3] Some items of data can be fuzzy in different ways, depending on the quality, the approximation, the probability or the multiplicity of values :

- subjectivity : it is connected with the quality of an object, for instance the specification of the nature of a fossil remain (attribute of the class *"SPECIMEN"*) can be qualified as doubtful ;

- approximation : it can relate to numeral or litteral data ; e.g. for the absolute age of a sample : "3,5 million years more or less 0,5 million years", or "about 3,5 m.y." ;

- probability : an item of data in paleontology is barely certain, it is more often probable with a possible percentage of preser-

vation of a fossil does not allow an absolutely certain determination ; nevertheless it is probably X, the probability of such a fact is about 70% ; the fuzzy item of data is then X 70% ; or there may be a strong probability for both the given specimens to belong to the same species ;

- multiplicity : it comes out when a set of values of indifferent cardinality is available rather than only one value, and when it is known that among these values only one is right but not which one among them. This must not be confused with the concept of table : there are not several numbered and simultaneously valid values, the case of multiplicity is often experienced in collections, for example when there are two different labels with contradictory information for only one specimen.

B. THE LACK OF DATA

The frequent case of lack of data must be considered, especially in the hierarchy. This is not in itself a fuzzy item of data, but it constitutes a gap in our knowledge. For example : though the class *"NAMED PLACE"* is a subclass of *"LOCAL SUBDIVISION"*, an object of *"NAMED PLACE"* may not have a predecessor if it is not known on which geographical territory the "named place" is situated (what can be done with such an information : "north side of ditch" without any other indication of localisation ?). So there are in the structure many "trees" of objects which are degenerated or which start at different levels to those of the classes composing the hierarchical structure to which they belong.

In the paleontological application, all the items of data are pratically incomplete. Approximation is implicit for geographical lacalisation in many cases. Subjectivity can spread on all that depends on ascertainments : either because these are very old or because there has been change in labels. We will note that a sample with several labels must not be dismissed or ignored. It must be treated like the others but with a detailed analysis of the information concerning it.
We will even state more generally, that all the specimens, too often left out in classified systems of collection curating because of the scarcity (even of the non existence) of data concerning them benefit, in this application, from a privileged treatment. This deliberate intention must avoid for certain samples potentially very useful to the progress of paleontology to be definitely forgotten about.

Nevertheless the reader may wonder what can be done with such doubtful samples. There are several answers :
- if a sample which is not exact, nor precise nor certain as for the data concerning it must be rejected, then no sample can be preserved ;
- statistically a number of such samples has a meaning even if each sample, separately considered can give no indication ;
- the treatment of such samples through computing methods can possibly allow to correct some of their uncertainties through research so oriented ;
- this sample is the scientific material to treat ; manually or with computers methods, it is the same material, it is a misconception to wish to adapt this material to any computing model, especially a data bank model ; this latter one must adapt to the material what ever its complexity ;
- this data represents the real world ; curators and users work on this real world and not on the characteristics of a computer.
The parts must not be reversed.

V. CONCLUSIONS

A SCHEME FOR A WELL ADAPTED TOOL AT THE DISPOSAL OF RESEARCHERS

This application to management and exploitation of paleontological collections is a tool with many uses which meets diversified objectives. It is no limited to a simple system of automatic documentation. The main objective is to allow researchers to have effective direct access to the collection material and to collect a maximum of information in and around the collection. Other applications already use statistic methods.

When this application will be functional, it constitutes a prospective study field which, overtaking the simple ambition to manager a stiff collection, must allow our knowledge in a field situated at the border between several Earth Sciences and Life Sciences disciplines to be extended.

REFERENCES

1. Bouillé F. (1977). Un modèle universel de banques de données, simultanément partageable, portable et répartie. Thèse de Doctorat ès Sciences Math., 550p. Paris.
2. Bouillé F., Pajaud D., Roulet M.-J. (1978). Paleontological data processing with an HBDS data bank at the Université Pierre et Marie Curie. In Pergamon Press (Ed.), Proceedings 6th CODATA Int.Conf., Santa-Flavia, Palerme (Sicile), 381-391.
3. Bouillé F., Pajaud D., Roulet M.-J. (1979). La banque de données des collections de paléontologie. Preprint Proceedings 1er Séminaire international "HBDS", Lisbonne (Portugal), 22p.
4. Pajaud D. (1977). Données actuelles sur le traitement de l'information des collections paléontologiques en France. Bulletin d'Information des Géologues du Bassin de Paris (Service Coll. Pal., doc. n°5), 14, 4, 57-73, 15 fig.

Daniel PAJAUD

Docteur ès Sciences naturelles (1967)

Maître-Assistant à l'Université Pierre et Marie Curie à Paris, Laboratoire de Paléontologie des Algues et Invertébrés

Conservateur des Collections de Paléontologie de l'Université Pierre et Marie Curie

Rédacteur de la Revue "Géologues" de l'Union Française des Géologues

Président de l'Association Nationale des Scientifiques pour l'Usage de la Langue Française