Geophysical Research Abstracts, Vol. 7, 03466, 2005
SRef-ID: 1607-7962/gra/EGU05-A-03466

# Data and databases for sedimentary geology and paleontology: The PaleoStrat approach for the CHRONOS System

**W.S. Snyder** (1), C. Cervato (2), and P.M. Sadler (3)

(1) Boise State University and PaleoStrat, Department of Geosciences, Boise, ID 83725, USA, (wsnyder@boisestate.edu), (2) Iowa State University and CHRONOS, Department of Geological and Atmospheric Sciences, 253 Science I, Ames, IA 50011, USA, (cinzia@iastate.edu), (3) University of California, Riverside, Department of Earth Sciences, 900 University Avenue, Riverside, CA, 92521, USA (peter.sadler@ucr.edu)

## Introduction

Layered sedimentary rocks are a vast, globally distributed, information repository for Earth history, spanning billions of years, that holds the answers to many scientific and societal questions. Countless research projects have moved much of this information to the printed literature where it is still expensive and time consuming to reorganize for each new question. The geoinformatics revolution is accelerating research by migrating information to more readily retrievable electronic formats. By augmenting community databases in sedimentary geology and paleobiology and giving them an unprecedented level of interoperabilty, the CHRONOS System will realize a virtual stratigraphic record - a means to boost the pace and enlarge the scope of integrative geoscience. The critical core of the effort is the PaleoStrat database system that offers a mechanism to handle a broad array of "time series" data; data tied to stratigraphic sections, whether that be from terrestrial surface sections or from wells and drill holes. PaleoStrat is a repository for the basic building blocks of the geologic record and encompasses, but steps beyond the repositories previously available only for the ocean drilling records. The interoperability has two key components. The first exploits software advances for translating between different data structures and terminologies. The second is CHRONOS's focus on evidence of geologic age, the universal means to combine stratigraphic information into a common time scale. Research projects continue to generate new information and mine the printed legacy data, but now they leave can their compilations within PaleoStrat or within reach of the CHRONOS System. This system maintains the virtual stratigraphic record and begins to solve the legacy data problem - those data that are most needed will be first mined and migrated to electronic formats and the CHRONOS System will assist in the migration. The goal of the CHRONOS System therefore, through its partners and its own direct products, is to provide the geoscience researcher with a comprehensive geoinformatics facility for sedimentary geology and paleontology (sensu lato). A fully implemented geoinformatics system must be comprehensive, modular, and extensible. It must meet the diverse needs of the research science and therefore encompass data and tools not typically considered core parts of sedimentary geology or paleontology. The modularity and extensibility is necessary to accommodate growth and evolution of thought in both the Earth and computer sciences. To better allow for this modularity and to more clearly communicate the needs of the geoscientists to the computer scientists, a conceptual model of the Earth sciences (in this case for a sedimentary-paleobiologic system, including time-series analysis and time scale/geologic age issues) must be developed that is suitable for designing the information system. This conceptual model should form the basis for an ontology (a formal or specific implementation of the conceptual model), and group and connect the relevant data and metadata into a logical data model that reflects this geologic framework. We outline here a basic conceptual model for sedimentary-ancient life systems. It is important to note that this is just a version, a step along the way of working with the community to better capture their needs in the system - hence, it should be read with the notion in mind that some of what is presented may, and indeed, should change over the years. The technical issues associated with database development are not addressed here. Furthermore, we have tapped many sources for this model, most importantly the sedimentary geology-paleontology community and co-workers on PaleoStrat and CHRONOS, but we have also utilized the North American Data Model for Geologic Maps (Association of American State Geologists - U.S. Geological Survey - Canadian Geological Survey), the Classification of Sediments and Sedimentary Rocks of the British Geological Survey, and other geoinformatics groups (e.g., PetDB, GEON, and NAVDAT in particular).

## Database Conceptual Model

The conceptual model, and its derivative, the data model, attempt to encompass the needs of: paleontology (micro and macro), biostratigraphy, paleobiology, lithostratigraphy, sequence stratigraphy, cyclostratigraphy, chemostratigraphy, chronostratigraphy, magnetostratigraphy, geochronology, paleogeography, basin analysis, tectonostratigraphy, and certain aspects of tectonics. The model must be able to accommodate data that is terrestrial or marine, from outcrops or wells/drill holes, analytical and descriptive, and accommodate storing photographs and diagrams as “data”. This does not have to be a “final” data model - indeed it cannot be so, simply because it must grow as the needs of the science grow. Nor does it have to encompass the entire breadth of the geosciences - it can be done in a “modular” way, provided that the linkages to other modules are articulated. The conceptual model should be simple enough to be useful for database and tool design, but sufficiently comprehensive to cover the above disciplines - in total or in part. It is important to encompass the needs of the more thematic questions, such as life evolution (origination, extinction, radiation), Earth’s chemical evolution, and deep-time sea level changes. It must support the needs of the research community in general, and support science research efforts such as EARTHTIME (www.earth-time.org), GeoSystems (www.geosystems.org), and ANDRILL (www.andrill.org). In general, such science themes and initiatives require only three components: 1) a database that includes all relevant data and metadata types, 2) a simple, but powerful way for the user to find the information they need, and 3) the analytical and assessment tools necessary to address the thematic science questions. Thus, the conceptual model must accommodate not only the physical, chemical, and biologic features of the rock record, but also the intended uses for these data. Again, because it is impossible to make a “final” statement on the latter, the system must be extensible. Furthermore, this model is put forward as an initial attempt that can only be improved as subdiscipline experts make corrections and add more detail. The basis for the conceptual framework presented here is:

Process proxies: Earth processes are recorded in the rock record as a variety of physical, chemical, and biological “signatures” that must be reflected in the data and metadata. Scale dependence: Within the database, data types, characteristics, resolutions, and relationships are scale dependent, both temporally and spatially. The type of data and metadata required by the domain science research is often dependent on the spatial and temporal scale of the investigation. Therefore, the geoinformatics system must accommodate the ability to “zoom-in” and “zoom-out” and gather only the data needed by the user for their specific investigation. Time series: The vertical succession of sedimentary strata record a time series of events and all objects and data collected through the stratigraphic stack reflect a time series of processes. This is complicated by gaps in the stratigraphic succession and when crustal deformation (fold, faults) disrupts this ideal layer-cake geometry. The rock record at any one site is therefore an incomplete proxy of time. One goal of sedimentary geology and paleontology is to correlate the geologic records of many sites, globally if possible, to better piece together a more complete geologic history and therefore to better understand the processes that have, and are shaping our planet Earth. Spatial patterns: there are lateral and vertical changes in physical, chemical, and biologic characteristics of strata, and the spatial distribution of an object can vary from microscopic to local to global. Spatial patterns reflect the fact that laterally, along any one time line of the rock record (think of a single stratigraphic plane or one card in the deck),objects change in a systematic, and potentially predicable ways - even if that change is complex or chaotic. That features and samples have spatial distributions should be obvious, but it is important to separate "known" global distributions from presumed ones. For example, sequence stratigraphy and cyclostratigraphy assume the global distribution of features related to sea level rise and fall and orbital climate forcing. The issue is to develop an independent data set that can test such assumptions rather than merely compile such interpretations; in the latter case it is far too easy to forget, and begin to think of interpretations as "data". Age, and physical and location: Every object collected or described from the rock record has an “age” of its origination and a “location”. “Age” can be expressed in terms of classic time scale, cyclostratigraphic, or magnetostratigraphic units, or as radiometric ages; they can be point values, or encompass intervals of the stratigraphic record. Ages can be denoted as "points" in time or "intervals" of time. We express location in terms of latitude and longitude (but these data can be input in a variety of formats). “Location” refers not only to location in present-day coordinates, but to paleo-coordinates within the original sedimentary basin and on the globe (i.e., paleogeographic coordinates). This is important for reconstructing the history of each basin, mountain system, continent, and the Earth. Location can be recorded in decimal degrees and a relative “temporal location” in stratigraphic meters from an arbitrary point within a stratigraphic section, well, or drill hole (these arbitrary points have latitude-longitude “locations”). Geologic objects and Attributes: Geologic objects represent physical and geophysical entities and can generally be divided into “regions”, “features”, "samples", and “subsamples”. Objects have “attributes” which encompass their physical, chemical, and biologic characteristics in addition to the “attributes” of age and location. These in turn can be measurements, analyses, or descriptions. These measurements, estimates, or analyses and descriptions can be direct, indirect, derived products, or interpretations.

## Translating the Concept into a Database

Working with the user community to insure that we capture all the data and metadata types they feel are necessary is key to developing a useful database. Such interaction is, of course, not sufficient, as the user often is not fully aware of all the needed data or the relationships among data and metadata types. Some are familiar with the difficulty of developing complex relational databases, but most are not. What is required are dedicated computer/information technology scientists to work with geoscientists to develop an exceptionably complex data model, and from that model, the database. We have attempted accomplish this via community interactions at professional meetings, workshops, and one-on-one associations. It is an iterative, on-going process that takes time and resources to accomplish. For example, the recent redesign and expansion of the PaleoStrat database by the CHRONOS development team was necessitated by this community interaction, but it took a full year to make the changes. Also, it is important to remember that even if the CHRONOS System provides a comprehensive sedimentary geology-paleontology database that this does not replace the need to continue and expand the federation of CHRONOS with other national and international databases.

## The Problem of Legacy Data

Legacy data is perhaps the most difficult data to capture in a database. Nevertheless, a motto for PaleoStrat could be: “Built for the future - but working to capture the past”. This reflects the importance of not only providing a mechanism for capturing and delivering data from future research, but that we must also capture the legacy data from past research. These are the data that, at least for the next decade, will continue to provide the basis for asking our most significant science questions, for framing what we think we do and do not understand about how the Earth works. During the transition from yesterday’s approach and tomorrow’s, the capture of these legacy data will be extremely difficult. Historically, we have been stuck in a system where all relevant data and metadata cannot be presented in a published article, and often not even found in “supplemental data” files that many publications offer. Hence, our legacy is one of incomplete data sets - and these missing data and metadata may not be recoverable for the vast majority of our past research efforts. This difficulty is often exacerbated because the researchers who generated the data are, as they must be, very busy doing new science, teaching, administrative work, and management. We also have a culture of not fully sharing our data - the “my data” syndrome. Therefore, because of the unavailability of complete data and metadata, the busy lives of researchers, and the historical culture of much of our geoscience, the capture of legacy data will be extremely difficult. Some other databases, such as PedDB (www.petdb.org), NAVDAT (www.navdat. geo.ku.edu), and GEOROC (georoc.mpch-mainz.gwdg.de/georoc/) (all igneous petrochemistry databases) have taken on this challenge by loading data and what metadata is available directly from the literature. This is a huge task for the principle investigators on these database projects, but it has proven fruitful. The challenge is to find, load and then quality check the data; often the author of the publication must be contact for additional information, which is all too frequently not supplied. Admittedly, these are more simple databases than that of PaleoStrat, in that they target a more restricted suite of data types, but the approach has proven workable, except perhaps for the lack of sufficient funding for the people who to load and quality check the data.

## Solution for legacy data in PaleoStrat:

The PaleoStrat learned this same lesson about legacy data. We had hoped, perhaps naively, that by providing web-based forms and Excel templates that we could encourage researchers to input their legacy data. We have been somewhat successful in this - but the results are inconsistent. This, coupled with the development of the new CHRONOS System has significantly increased the potential size of our collective data holdings. The capture of legacy data will now be addressed by continuing to offer web-based forms, standard Excel templates, but also by setting up a data loading system comprised of a supervisor and undergraduate geoscience students. The goal is to make it as easy as possible for scientists to contribute their legacy data. Some may take advantage of our online data forms, others of Excel templates, but many others will only have random Excel files, old, home-grown databases (e.g., in Access), or merely paper copies of published articles and paper data in their file cabinets. We will work with all interested scientists, but we also realize that we will have to prioritize our efforts. Finally, we are enthusiastic partners of the SESAR project (www.geosamples.org) to develop an international system for a unique sample identifier (IGSN - International Geologic Sample Number), which will greatly improve the quality of captured legacy data.