The ESP Journal

Extraordinary insight into today's education information topics

Re-Defining Data

Glynn D. Ligon, Ph.D., ESP Solutions Group, Inc.

ESP Solutions Group

Table of Contents

Contexts

Context: This paper is written within the domain of learning, the kingdom of education, and the phylum of decision making. Oh, yes, we'll describe a taxonomy to make the phylum reference make sense and to make decision making rise to the level of highest importance.

Audience: The primary audience is everyone who makes a decision and takes some action related to learning/education.

Level of Academic Scholarship: This paper is not a journal article. The advice and insights are serious. However, the presentation is informal.

How to Use This Paper: This is intended mainly as a reference document. Reading it from start to finish might be a choppy experience.

Characteristics of Data

In 2006, I wrote the original paper "Defining Data[1]." Surprisingly little had to be updated after 18 years. However, I did want to emphasize more that the primary purpose of data is informing decisions.

I did add the Learning/Education Classification Taxonomy to describe the "Kingdom of Data." There're also now the PII Charts. Confidentiality and protection of personally identifiable information are just too important not to have a couple of graphics.

Let's begin our re-defining of data with the new Learning/Education Classification Taxonomy.

Real-time data work for making timely decisions by teachers and other educators. Validated, official statistics work for state reporting (and ED*Facts*), longitudinal analytics, growth models, research, etc. Does your state's SLDS[2] do both well? Is your data model founded upon providing data for decision making? Maybe your data model has been created to describe whatever data are being collected now—no matter how those data were defined, compiled, and offered up for decision making.

ESP has documented over 50 adjectives describing data (formative, summative, imputed, personally identifiable, discrete, categorical, etc.). Bottom line, we have to match the nature of our data with the intended use of them to ensure valid decisions.

What better way to examine our education data than to compare ourselves to the classification of the animal kingdom--us.

The Kingdom of Education Data

Education IT pros are organizing our data using data models. Let's step back from the details of those data models for a moment. For example, the Common Education Data Standard (CEDS) has worked diligently to create nine domains with corresponding entities, categories, and elements for education agencies to model their systems around. However, I just reviewed their documentation, again. https://ceds.ed.gov/publications.aspx Something's missing.

The workmanship is fine. The impression given off is that a committee began right in the middle of a problem and is methodically working its way outward.
That's probably how most, if not all, large data models develop.

Let's go even broader.

For centuries, scientists have been arguing about classification of animals and plants--taxonomic ranking. Are there eight or nine classifications, multiple subclasses, and ancestral or hereditary hierarchies? The competitions for classification dominance predate Aristotle. This is the style of intellectual debate we should be having over education data. Not just whether there should be a category for PK12W rather than just K12.

So, here we go. I propose a rethinking of education data into a classification system that parallels that of the animal world. To begin this process, we must have some fundamental understandings.

1. We collect data about learning within the education universe to inform decisions.
2. The decisions to be made should determine the data we collect, manage, and provide to the decision makers.
3. Everything else is process[4].

Our Human Classification

The most agreed-upon modern human classification has eight levels. These are typically shown as an inverted pyramid.

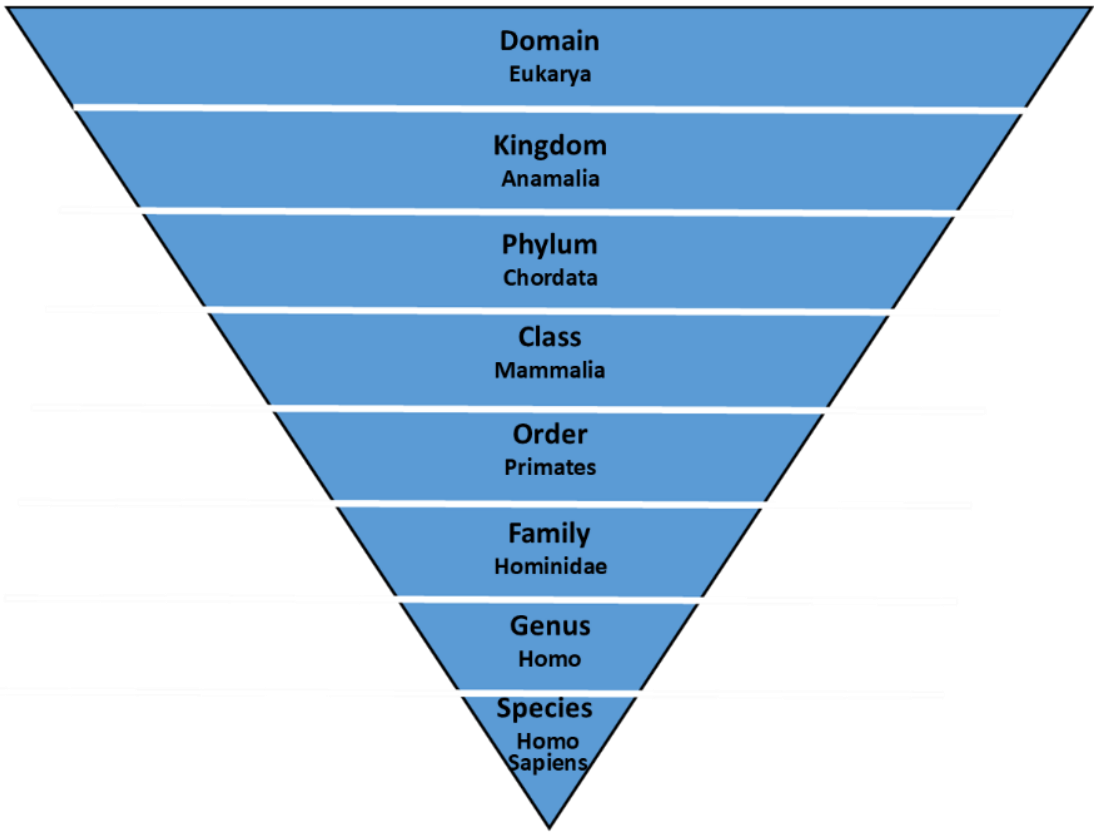

Figure 1: Classification System for the Animal Kingdom Shown for Homo Sapiens (Us)

Our Education Data Classification

Our vision for education data is quite different. There may be fewer domains and room for many more data elements, which this inverted taxonomy allows. Our pyramid is not so tipsy.

Our domain is Learning. Others might be Commerce, Governance, and Technology.

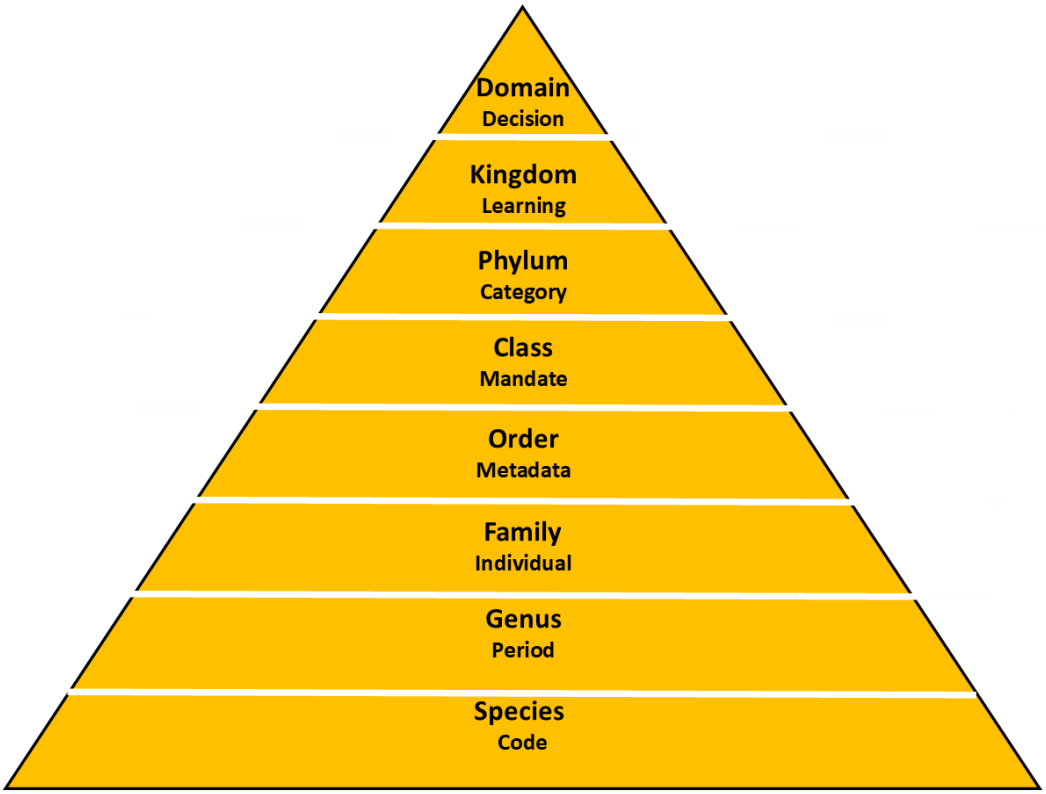

Figure 2: Classification System for Education Data

ESP Solutions Group has worked with most state education agencies and the U.S. Department of Education to create data dictionaries. We have compiled quite a collection of adjectives describing data. These adjectives are significant because each denotes a use, condition, purpose, readiness, or other state of data that is crucial for a data manager or decision maker to understand.

Classification Topology: Humans & Learning				
Classification (Taxonomy)	**Humans**	**Learning/ Education Data**	**Definition of Classification in Learning/Education:** Data Adjectives	**CEDS Classification**
Domain	Eukarya	Learning	**Decision Making:** Actionable	
Kingdom	Anamalia	Education	**Learning/Teaching:** Education, Academic, Research	
Phylum	Chordata	Mandate	**Decision Making/Purpose:** Funding, Input, Out-come, Intervention	
Class	Mammalia	Category	**Category:** Context, Subject	
Order	Primates	Metadata	**Content/Operational Definitions:** Deidentified, Transactional, Performance, Snapshot, Formative, Summative, Outcome, Process, Diagnostic, Historical, Longitudinal, Predictive, Lagging, Leading, Max-Yield, Meta, Conflicting, Extract	Domains
Family	Hominidae	Individual	**Entity (Individual, Person, Object, Owner):** Personally Identifiable, Confidential, Deidentified, Role, Demographic	Entities, Categories
Genus	Homo	Period	**Periodicity/Quality:** Realtime, Point-in-Time, Clean, Quality, Raw, Derived, Extract, Backup, Missing, Imputed, Obfuscated	
Species	Homo Sapiens	Code	**Type (Text, Code):** Backup, Continuous, Discrete, Categorical, Interval, Electronic	Elements

Figure 3: Education Data Classification Topology

In Figure 3, some of these adjectives for data elements are displayed in our newly created taxonomy for education data.

"In Attendance" Data Represented in the Learning Taxonomy			
Generic Classification (Taxonomy)	**Classification for Learning Data**	**Definition of Classification in Learning**	**Sample Data Element: "In Attendance"** (Student Present or Absent on January 31, 2024)
Domain	Learning	Decision Making	Decision Making
Kingdom	Education	Learning/Teaching	Learning/Teaching
Phylum	Mandate	Category	Management
Class	Category	Mandate/Purpose	Funding, Intervention, Counseling
Order	Metadata	Content/Operational Definitions	Operational Definition
Family	Individual	Entity (Individual, Person, Object, Owner)	Student, Class, Grade, School, District, State
Genus	Period	Periodicity	Single Date (Day, Month, Year)
Species	Code	Type (Text, Code)	Code (Present; Absent/ Excused/Unexcused; Tardy)

Figure 4: "In Attendance" Data Represented in the Learning Taxonomy

Let's look at student attendance as an example. The presence or absence by a student on a given date is a common data element in a data model and a database. Figure 4 aligns the data element "In Attendance" to each level of the Learning/Education Classification.

Comparison to CEDS

I mentioned earlier that CEDS's data model used domains, entities, categories, and elements. See Figure 3. Despite some overlap in terminology, these groupings do not span the range of our new Education Data Classification Topology. This perspective is consistent with the opinion that CEDS, as a data model, describes data as objects. To us, data are far more. Data are essential components of decisions. Data are building blocks for decision makers to follow their plan for action.

Yes, we need a data model. However, that data model must be within the much broader and more significant context of a comprehensive data classification system. That is how we maintain our focus on the true purpose for data—decision making.

PII Charts

Technically, we're going to use Venn diagrams as our PII charts to discuss personally identifiable information (PII).

Figure 5 simplifies PII into its four categories as defined by FERPA[3].

The first direct data elements identify a student. The term includes, but is not limited to—

1. The student's name,
2. The name of the student's parent or other family members,
3. The address of the student or student's family, and
4. A personal identifier, such as the student's Social Security number, student number, or biometric record.

The indirect data elements allow a person to infer who a student is. These might be—

1. Date of birth,
2. Place of birth, and
3. Mother's maiden name.

The third is other information defined by FERPA as—

> Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty.

The fourth category is rather general. In fact, in this one, the person already knows the identity of the student and is seeking other information.

> Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates.

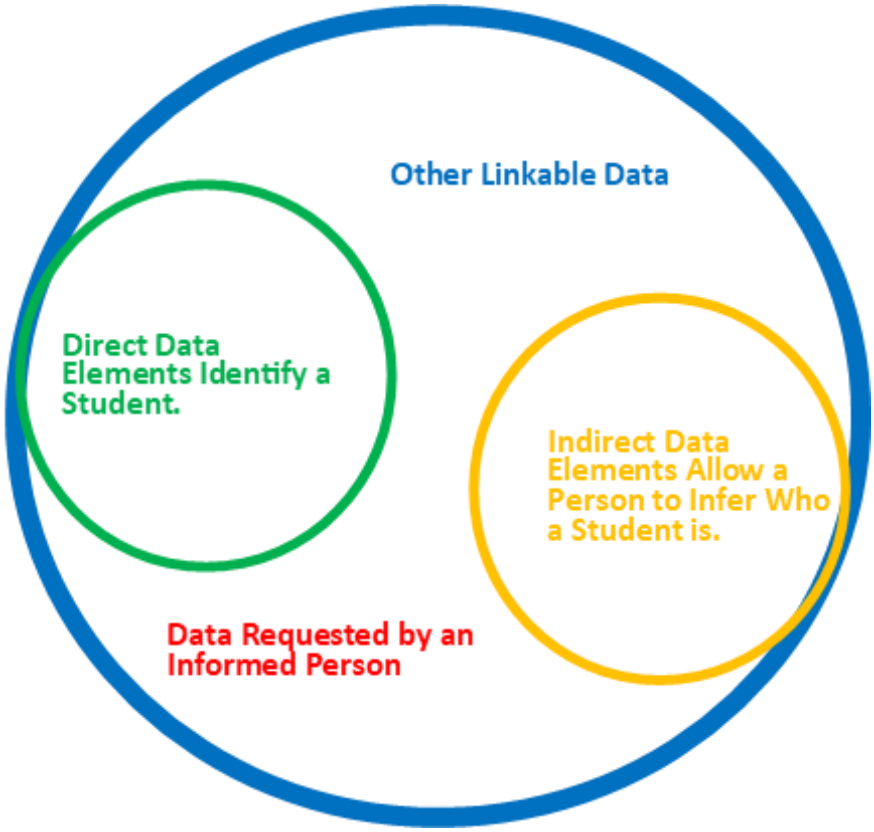

Figure 5: FERPA's Four Categories of Personally Identifiable Information (PII)

With that specificity, our next chart is Figure 6. This shows how PII is shared among virtually all data systems within an education agency. This results in the responsibility for understanding FERPA, a state's interpretation if it into its laws, and all local regulations is an agency-wide responsibility.

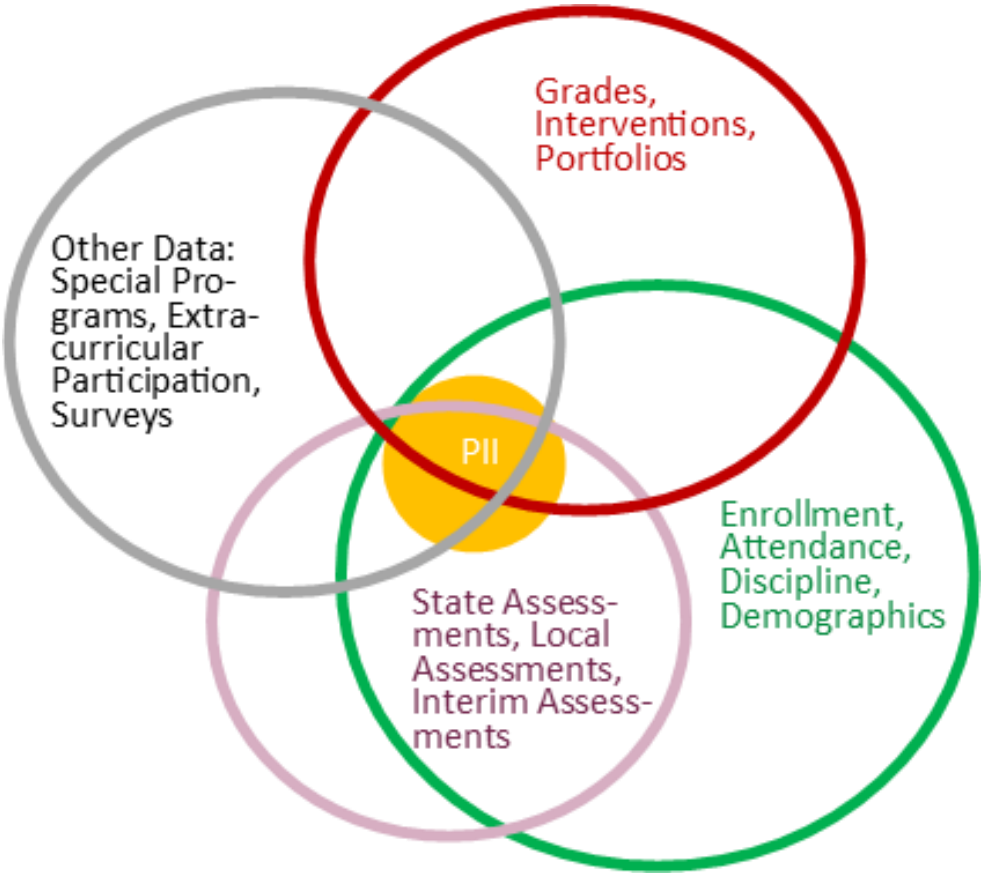

Figure 6: Shared Responsibility for PII Across Agency Data Systems

Defining the Educational Process with Data

What are the characteristics of the data that define the educational process? You may have thought this paper would be about how we define data elements. To a great extent it is, but more importantly, a discussion of the nature of the data that we use to define (to describe) our schools, student academic progress, and accountability indicators.

Hold off for now on the traditional, bland definition of data.

> *Data: noun. 1. plural of datum. 2. individual facts, statistics, or items of information.*

I want to define data in today's context, by the characteristics that data must have for us to incorporate them into our data driven decision making processes. In other words, what are the characteristics of a data element that qualify it to be included in one of our mission-critical information systems? If data do not measure up to these criteria, they are noise. Worse, they are in the way.

Today's education data must have these attributes.

1. Defined adequately such that the providers, processors, and users of the data all have the same understanding of what is being described, measured, or reported
    - A metadata dictionary, a user guide, and technical documentation team together to provide clear and precise definitions and characteristics
2. Aligned to an open standard such that when the data are exchanged between information systems, both the source and destination software applications correctly interpret the values/content
    - Individual applications conform to interoperability standards (e.g., SIF, Ed-Fi, CEDS, etc.).
    - Open standards allow data exchanges beyond the scope of a single vendor's reach.
3. Specified in their periodicity such that the providers, processors, and users of the data know the time period represented by the data and the collection and reporting schedule for the data
    - The metadata dictionary and user guide specify the time period from which the data are collected and reported.
    - A data collection and reporting calendar document when the data are available for use.
4. Collected at a level of detail that allows analyses, queries, and reports aligned with the questions being asked by decision makers
    - The granularity of the data allows for re-analysis and disaggregation to meet changing decision needs without re-collecting data.

5. Collected because they are needed for a specified purpose and not available from another source
    - An organization's overall data management process ensures that only useful data are collected and that they are collected once and shared for many uses.
    - Efficient collection save data providers time and effort for their primary jobs.

6. Stored digitally
    - To fit into today's information systems, anything to be saved and accessed later must be digital. By this definition, if information items are not digital, then they are not data—for our purposes.
    - No value judgment is being made, just the practical reality that our new information systems process digital data.
    - Practically everything can be converted to a digital image these days—chemicals, classical paintings, music, colors, etc.

7. Stored in a schema that optimizes access by a user versus not efficient use of storage space
    - Recall when best practice mandated that our databases be normalized—every datum stored one time in such a logical way as to eliminate all redundancy?
    - Now the emphasis rightly so is on speed of access. We store data so we can find them and use them. Who cares if that means having the same data element in the database a dozen times?
    - A single data warehouse is not the most efficient way to manage all of an education agency's data. Data consolidation and access are complex challenges that should be driven by the use of the data rather than a trendy data warehouse solution.

8. Validated against data rules that ensure compliance with standards, definitions, database formats, etc.
    - Definitional data rules ensure validity.
    - Format data rules ensure interoperability and access.
    - Relational data rules ensure that the data make sense in terms of the other data within the system.

9. Related to other data that together provide the insights into what is really happening with students in our systems
    - Disaggregating data for subgroups as required by most accountability systems means we must be able to put the same student in multiple groups dependent upon that student's characteristics.
    - Growth, value-added, longitudinal, and other research and accountability models require linking across years, assessments, school characteristics, and student characteristics.
    - Benchmarking and other comparative processes typically call upon multiple indicators across multiple entities.
    - Calculating rates (dropout, attendance, graduation, retention, passing, discipline, etc.) requires both a numerator and a denominator with the appropriate periodicity.

We have become very demanding of our data. The best emerging data management, analysis, and reporting systems are being developed by educators and vendors who appreciate the fact that education data are different from traditional business data. An intimate knowledge of how teachers and students interact, how schools and districts operate, and how states fund and support them is crucial today.

Recurring Data Themes

The following is a somewhat irreverent review of persistent issues related to data. A few of these are trivial, but interesting. A few are core to our understanding of data and maximizing their use.

Significance of Data—A.K.A. Statistics

How reliable or statistically significant is a statistic? How much trust should we place in one versus another? For example, if an adequate yearly progress report says that a student subgroup had 75% proficient on a state assessment, how reliable is that statistic? Statistical significance tests estimate the likely swing in that statistic if multiple measurements were to be made. See ESP's Optimal Reference Guide, *Confidentiality and Reliability Rules for Reporting Education Data*[5], for an in-depth analysis of these issues.

Researchers know that the number of students in a group determines the reliability, and if more than one group is being compared, that whether or not they are of the same size makes a difference. Generally, the larger the groups, the more reliable the differences between them. To make interpretation of statements of statistical significance clear to the reader, I suggest that we write the word "significant" to communicate the group sizes being compared.

- SigNificaNt: Two groups are large and equivalent in size.
- SignificaNt: Two groups are unequal in size, one being large and the other small.
- Significant: Two groups are small but of equal size.

A reader would know immediately that the difference between the two groups must be large if "significant" is used. The difference between the two groups could be very small when "sigNificaNt" is used.

Terminology

I do believe that the word data is plural—and datum is singular. However, common practice accepts data as a collective noun, thus singular as well. So if staff, jury, and other nouns can be considered singular or plural dependent upon context, then so can data. I shy away from this logic because then we would need to determine if the data are acting as a group or as individuals within a group to know if the verb should be singular or plural (e.g., the jury are of different opinions; the jury is unanimous.) I am not ready to designate data as singular, so my concession in this paper will be to refer to a data element when a single datum is referenced.

Data are not Just Numbers Anymore

In today's world, data describe not only the number of students in attendance but also an individual's performance-art project on DVD. Portfolios, body-of-evidence systems, qualitative assessments, PowerPoint shows, and photographs are all data. For example, the spreadsheet, originally developed to do the mathematics that accountants perform, has evolved to produce graphics and hyperlink to videos.

Data are All Numbers

However, even though a very subjective or visual construct is being described, the data used to document that description is now digital—zeroes and ones. Digital representation of data is now a necessity. Storage, Internet transmissions, burning a CD/DVD, etc., all require a digital coding.

Storage Capacity is No Longer an Issue

This is one of the most significant advancements related to data. We can now be data packrats without guilt. In fact almost all the other issues touched on in this paper have been influenced by the availability of cheap data storage.

Transmission Speed is No Longer an Issue

The improvements in transmission speeds have also exacerbated the proliferation of data. Enormous compressed files move efficiently between school districts and state education agencies these days. Files are shared without much thought given to transmission time. A r e s ystems still slowed down by too many concurrent users?

The Data Quantity Conundrum

Where do we draw the line between all the data that can be collected and all the data that are fit to be collected? That is where the nine criteria stated at the beginning of this paper provide guidance.

Granularity

The advancements in data storage have a fantastic benefit for our data. We can now store data at whatever level of granularity that is appropriate. If you are not convinced of the sigNificaNce of granularity, here is a comparison of a couple of states after the No Child Left Behind Act was passed.

- State 1 with an individual student record system that allows calculation of student subgroups as defined by NCLB: New calculations were required to match the disaggregation rules of NCLB. No new data had to be collected from the schools.
- State 2 with aggregate statistics collected for a minimum of subgroups as required for state funding: New statistics had to be calculated by districts and reported in aggregate form to the state.

Since the passage of NCLB and the resultant state accountability systems, states without an individual student record system have made transition. When subgroup definitions change, they are ready.

The Salsa Scale

After Vince Paredes (Vice President of Research and Development for ESP Solutions Group) championed the notion of enhancing the granularity of data within information systems, Barbara Clements (Chief Standards Officer, National Transcript Center, ESP Solutions Group) and I were having lunch debriefing from an NCES conference. Her lunch was to have included pico de gallo, a chunky mixture of vegetables and peppers. What she got was picante sauce, smaller bits in an almost liquid state. The ESP Salsa Scale was born. As illustrated in Figure 7, the granularity of pico de gallo allows analysis of the contents, whereas, the blending of ingredients in picante sauce hides the detail.

The point of the salsa scale is that the more we blend our data and lose granularity, the fewer options we have to disaggregate the parts and understand what our students are really like. Barbara and I still debate whether ketchup or V-8 juice is the lowest end of the Salsa Scale, but we both agree we would not bother dipping a chip into either one to examine the contents.

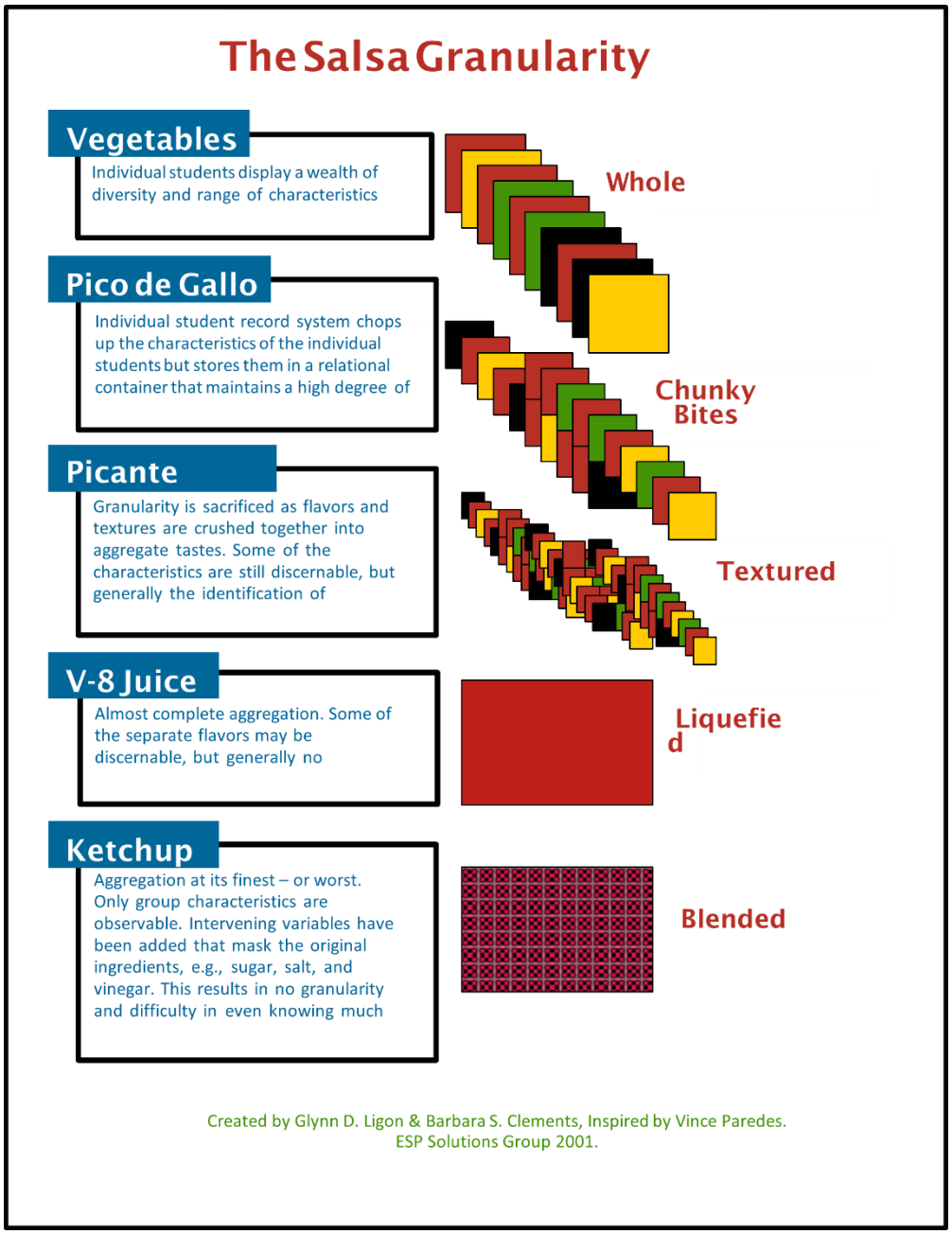

Figure 7: The Salsa Granularity Scale

Basic Raw Data Elements vs. Derived Data Elements

This debate will continue. The two sides are represented by these perspectives.

1. Store only the basic raw data elements because the derived statistics can always be calculated on-demand. In fact, if the derivation formula changes, then the old statistics do not need to be replaced while everyone worries that someone will use the old statistics rather than the new ones.
2. Store both the basic raw data elements and the derived statistics to ensure that derivations are correct and to speed processing time.

To a large degree, the first position, storing only the basic raw data elements, is a carryover from the old days of limited storage space. More and more decisions are being made to store both raw basic data elements and derived elements—for efficiency of access.

Data are not Facts

Upon reflection, we find that the word data may carry with it a connotation that is not quite deserved. Data are often taken as facts. In fact, the dictionary definition calls them facts. As I read the daily newspaper, I am often struck by what passes as a fact. Reporters print a quote from a key source and if what is reported later proves to be incorrect, the reporters' defense is that they were merely reporting what they were told.

This analogy is too true in education. What gets reported becomes a fact, an official statistic, whether or not it is accurate. So we should always treat data as reported information and make an independent determination of whether or not they are really factual.

People and Data

Data quality has more to do with people than with data. We are moving toward an environment in which unobtrusive measures are recorded by software applications as we do our normal work, rather than asking people to stop their work to fill out reports. Even with unobtrusive data collections, people provide data. Dependent upon how well-trained, motivated, conscientious, skilled, and busy they are, we get quality data. Our automated systems faithfully perpetuate the errors that people make. Interoperability among software applications ensures that errors are shared quickly and efficiently.

Warning Labels

Taking the food package labeling analogy a step further, what if warning labels were required? For example:

Warning: Studies have shown that dropout rates are not comparable across states.

Warning: Studies have shown that "proficiency" is more difficult to achieve on some state assessments than on others.

Warning: These data were reported by busy people with other priorities.

Warning: These are the data we could get.

Warning: NCES national averages are typically three school years old.

The Non-Proliferation Treatise for Education Data

We need to endorse a treatise that the proliferation of education data threatens the data quality and support of the data that have maximum use for educators. At the point our automated systems with virtually unlimited storage and supersonic transmission speeds have tempted us to collect all the data we can possibly envision, someone will need to champion a house cleaning.

<u>Final List of Essential Reminders about Data</u>

Please keep in mind that this Education Data Classification Topology has not been vetted by our professional peers. This is not presented as anything more than a discussion starter. No wait, this is really more than that. This is a wake-up call to those working with data models. Please…

1. Remember, your data have to be organized around the decisions they are funded to inform.

Oh. There really aren't any other points to be made.

References

1. Defining Data, ESP Optimal Reference Guide, 2006, Glynn D. Ligon
   http://www.espsolutionsgroup.com/espweb/assets/files/ESP_Defining_Data_ORG-1.pdf

2. SLDS State Longitudinal Data System

   Systems developed by state education agencies with grants from the National Center for Education Statistics, Institute of Education Sciences. https://nces.ed.gov/programs/slds

3. Family Educational Rights and Privacy Act (FERPA) https://www.ecfr.gov/current/title-34/subtitle-A/part-99

4. American Productivity and Quality Center (APQC) https://www.apqc.org/

   Jack Grayson, Ph.D. founded APQC based upon the premise that successful companies achieve productivity through effective processes. The Malcom Baldrige National Quality Awards given by the National Institute of Standards and Technology (NIST), an agency of the United States Department of Commerce, are his legacy.

5. Confidentiality and Reliability Rules for Reporting Education Data, 2008, Glynn D. Ligon, Ph.D. and Barbara S. Clements, Ph.D.
   http://www.espsolutionsgroup.com/espweb/assets/files/ESP_Confidentiality_Reliability_ORG.pdf

About ESP Solutions Group

ESP Solutions Group provides its clients with *Extraordinary Insight™* into P20W education data systems and analytics. Our team is comprised of industry experts who pioneered the concept of "data-driven decision making" and now help optimize the management of our clients' state and local education agencies' information systems.

ESP personnel have advised school districts, all state education agencies, and the U.S. Department of Education on the practice of P20W data management. We are regarded as leading experts in understanding the data and technology implications of ESSA, SIF, Ed-Fi, ED*Facts*, CEDS, state reporting, metadata standards, data governance, data visualizations, and emerging issues.

Dozens of education agencies have hired ESP to design and build their longitudinal data systems, state and federal reporting systems, metadata dictionaries, evaluation/assessment programs, and data management/analysis and visualization systems.

To learn how ESP can give your agency *Extraordinary Insight™* into your P20W education data, contact us at (512) 879-5300 or info@espsg.com.

This document is part of *The ESP Journal* Series, designed to help decision makers analyze, manage, and share data in the 21st Century.

ESP Solutions Group

(512) 879-5300

www.espsolutionsgroup.com