

*The Optimal Reference Guide:*  
**Defining Data**

*Extraordinary insight into  
today's education topics*

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



**ESP Solutions Group**



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, this is 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 be:

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).
  - 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.



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



*Data are defined by their Periodicity.*

- The time period represented.
- The schedule for collection.

## Characteristics of Data continued

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.
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.
  - 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, 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 schools.
  - Disaggregating data for subgroups as required by the No Child Left Behind Act 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

How reliable or statistically significant is a datum? 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*, 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.

 **ESP Insight**  
*Because storage capacity is no longer an issue, we can now be data packrats without guilt.*

### Transmission Speed is an Issue

Of arguably equal importance to storage capacity, the improvements in transmission speeds have enabled 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. Systems are still brought down by too many concurrent users, but scheduling and other strategies are helping. For most education agencies, transmission speeds remain an issue or at least a nagging doubt as the volume of data increases.


### 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.

 **ESP Insight**  
*Since the passage of NCLB, states without an individual student record system have made or have planned the transition.*

Since the passage of NCLB, states without an individual student record system have made or have planned the transition. When subgroup definitions change, they will be 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 on the following page, 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.



## The Salsa Granularity Scale

### Vegetables

Individual students display a wealth of diversity and range of characteristics that describe them as unique persons.

### Pico de Gallo

Individual student record system chops up the characteristics of the individual students but stores them in a relational container that maintains a high degree of granularity.

### Picante Sauce

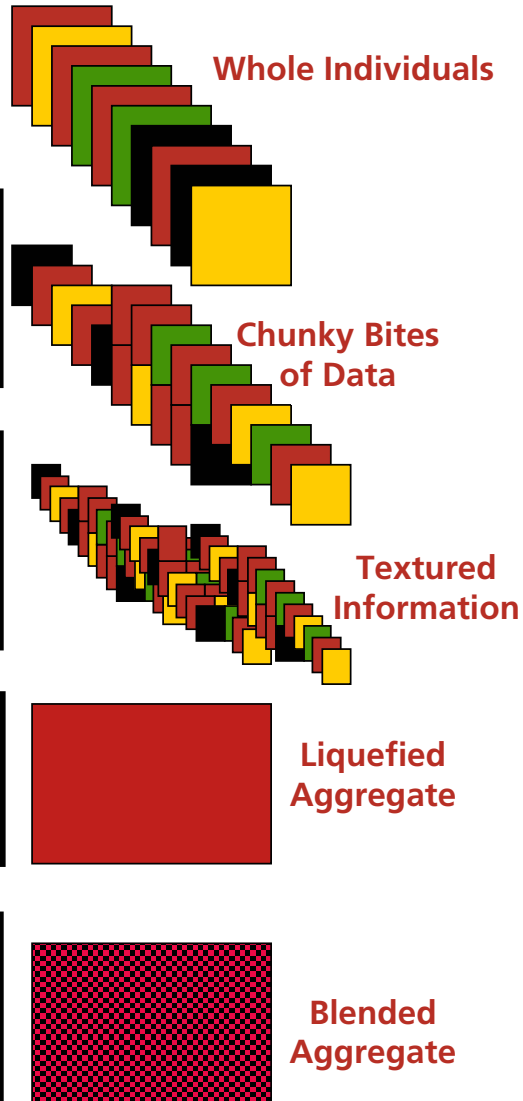
Granularity is sacrificed as flavors and textures are crushed together into aggregate tastes. Some of the characteristics are still discernable, but generally the identification of individuals is masked.

### V-8 Juice

Almost complete aggregation. Some of the separate flavors may be discernable, but generally no granularity is left.

### Ketchup

Aggregation at its finest – or worst. Only group characteristics are observable. Intervening variables have been added that mask the original ingredients, e.g., sugar, salt, and vinegar. This results in no granularity and difficulty in even knowing much about the original ingredients.



Created by Glynn D. Ligon & Barbara S. Clements, Inspired by Vince Paredes.  
ESP Solutions Group 2001.

### 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.




### Shelf Life

Data elements should have expiration dates on them. "These data are best if used before January 1, 2008." Why not? At the least, the data elements need to be documented in a metadata dictionary that tracks changes in definitions, codes, and quality that need to be considered.

### 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.


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

### 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 super sonic transmission speeds have trapped us into collecting all the data we can possible envision, someone will need to champion a house cleaning.

### Predictions about Data Proliferation

1. The amount of data collected will continue to grow in the short term, at the expense of teachers and other educators trying to get their work done.
2. Just because education data systems can collect more data, faster than ever, they will.
3. By the time someone questions how much useless data are filling data warehouses; those data warehouses will have reached an incredible 10% of capacity.
4. At that time, the data warehouse advocates will have imported only 25% of all the data they envision.
5. The drive to fill the data warehouses will subside when evaluations of the use of the data show that 80% rule applies (80% of the data in the data warehouse is easy to get and import, and requires 20% of the overall effort to fill the data warehouse. However, 80% of the effort is required to get the 20% of the data that are really useful.)
6. As query tools are opened up to a broad range of users, two trends will become evident. Very few people earn the title of user. A very small number of users generate a very large quantity of spurious analyses that challenge the professional analysts to verify their conclusions.

 **ESP Insight**  
*By the time someone questions how much useless data are filling data warehouses; those data warehouses will have reached an incredible 10% of capacity.*



**About ESP Solutions Group**

ESP Solutions Group provides its clients with *Extraordinary Insight*™ into K-12 education data systems and psychometrics. 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.

ESP personnel have advised school districts, all 52 state education agencies, and the U.S. Department of Education on the practice of K-12 school data management. We are regarded as leading experts in understanding the data and technology implications of the **No Child Left Behind Act (NCLB)**, **Education Data Exchange Network (EDEN)**, and the **Schools Interoperability Framework (SIF)**.

Dozens of education agencies have hired ESP to design and build their student record collection systems, federal reporting systems, student identifier systems, data dictionaries, evaluation/assessment programs and data management/analysis systems.

To learn how ESP can give your agency *Extraordinary Insight*™ into your K-12 education data, contact Greg Nadeau at (781) 370-1017 or [gnadeau@espsg.com](mailto:gnadeau@espsg.com).

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

*Defining Data*, Copyright © 2006 by ESP Solutions Group. All rights reserved. No part of this paper shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher.



**ESP Solutions Group**

**(512) 458-8364**

**[www.espsolutionsgroup.com](http://www.espsolutionsgroup.com)**

**Austin • Boston • Washington DC**