Rendering Students Legible: Translation Regimes and Student Identity in Higher Educational Data Systems

Presented at the Annual Meeting of the
Western Political Science Association
Seattle, Washington
April 17, 2014.

This paper takes a constructivist approach to understanding data in public-sector information systems, focusing on student data systems in higher education. It establishes the translation regime as a mechanism by which the social construction of data takes place, and suggests that translation regimes should be viewed as political structures rather than technical ones. Data exists because organizations such as universities or states have a need to make the domains in which they act legible. Doing so, however, requires some process that narrows the many possible representations of a given state of the world to a single data state. This process is carried out within translation regimes: systems of technical rules and social practices that establish a one-to-one correspondence between a given state of the world and a data state. The technical structures of a relational database, such as tables, functions, business rules, and queries translate states of the world into data states based on standards established by social structures such as cultures, states, and organizations. These regimes operate in a non-neutral fashion, carrying out a set of characteristic translations that favor certain groups over others. As such, information systems design is a political act, among other things shaping representation, asserting and protecting interests, and constructing normalized and deviant identities. Because these political acts are carried out through the technical structure of the translation regime, they appear as technical outcomes, making it more difficult to challenge them.

Jeffrey Alan Johnson, Ph.D.

Assistant Program Director,
Institutional Effectiveness and Planning
Utah Valley University

800 West University Parkway • Orem, Utah 84058
+1.801.863.8993 • jeffrey.johnson@uvu.edu • @the_other_jeff
http://uvu.edu/insteffect • http://the-other-jeff.blogspot.com

Rendering Students Legible: Translation Regimes in Higher Educational Data Systems

1 INTRODUCTION

Data, it seems, is to be the savior of the 21st Century. Whether in business, government, or higher education, pressures toward “data-driven” or “evidence-based” decisions are ubiquitous, promising more insight, more efficiency, and better outcomes than was previously possible. Implicit in this view, however, is a scientifically realist view of data: data can save us because it is an objective representation of observed reality that can thus transcend politics to bring organizations to the correct decision. If this view of data is incorrect, however, the edifice that legitimizes data becomes far less stable. If data is a social construct, requiring acts of choice and interpretation in its creation, then it becomes political, it’s power masked behind its false realism.

This paper takes the latter approach to understanding data. It establishes the translation regime as a mechanism by which the social construction of data takes place, and suggests that translation regimes should be viewed as political structures rather than technical ones. Data exists because organizations such as universities or states have a need to make the domains in which they act legible. Doing so, however, requires some process that narrows the many possible representations of a given state of the world to a single data state. This process is carried out within translation regimes: systems of technical rules and social practices that establish a one-to-one correspondence between a given state of the world and a data state. The technical structures of a relational database, such as tables, functions, business rules, and queries translate states of the world into data states based on standards established by social

structures such as cultures, states, and organizations. These regimes operate in a non-neutral fashion, carrying out a set of characteristic translations that favor certain groups over others. As such, information systems design is a political act, among other things shaping representation, asserting and protecting interests, and constructing normalized and deviant identities. Because these political acts are carried out through the technical structure of the translation regime, they appear as technical outcomes, making it more difficult to challenge them.

2 METHODOLOGY

To explore the nature of translation regimes, I examine the data system in place at Utah Valley University (UVU) while I worked as a Senior Research Analyst in its Institutional Research & Information office from 2009-2013.[1] That experience involved extensive work in data extraction and limited database design and administration, primarily in the Banner ODS database. This is supplemented by narrative analysis of the Structured Query Language (SQL) implementing the data systems and the data standards established by the federal Integrated Post-secondary Education Data System (IPEDS) and the Utah System of Higher Education (USHE) reporting processes, and occasionally by analysis of online interviews with eight UVU students regarding their own perceptions of their social identities conducted as a (highly unsatisfactory) pilot project for a larger study.

UVU's data backbone is the Ellucian Banner relational database running on an Oracle 10g database server.[2] Banner consists of normalized set several thousand data tables

[1] The analysis presented here is strictly my own, and should not be taken as representing in any way the policy of Utah Valley University or the views of its leadership.

[2] As a full review of database structure and operation is beyond the scope of practicality in a paper, tedious for those already familiar with them, and redundant give the many excellent sources available, this discussion presumes a basic, non-technical understanding of databases. I have aimed to provide enough background to understand the points in my argument in ways that do not overly burden those unfamiliar with databases with technical knowledge but are still

managing student and administrative data and optimized for Online Transactional Processing (OLTP), locally referred to as “Prod” (a reference to it as the production database). The bulk of institutional data analysis is performed using the Banner Operational Data Store (ODS), which consists of a denormalized set of fewer but much larger tables optimized for Online Analytical Processing (OLAP). The data contained in the ODS is either identical to or derived from that in Prod. Both databases are extensively customized for UVU. Prod also connects to several other data systems, including the Wolverine Track advising information system, Ellucian Student Success CRM, and the Canvas learning management system.

Most government reporting comes from three customized relational tables. One table, referred to locally as STUDENT,[3] contains information that is constant about individual students across courses within a term such as demographics, contact information, or overall academic characteristics. The second table, COURSE, contains information that is constant across all students in a section for a term. The final table, STUDENT_COURSE, contains information specific to a student within a specific course such as course grade or (since some courses can award variable credit) credits attempted. Using appropriate joins, STUDENT, COURSE, and STUDENT_COURSE can provide most of the information that the institution would need to understand its students and academic offerings. For example, joining STUDENT and STUDENT_COURSE would allow the institution to determine the distribution of courses taken by major and gender. STUDENT_COURSE would identify the courses taken by each student; STUDENT would provide the major and gender information. Each table is a “live” data table,

recognizable to technical specialists. I apologize to readers of both sorts to the extent that I haven’t succeeded in that.

[3] Table and field names will be indicated in capital letters, with TABLE_NAMES in Roman typeface and *FIELD_NAMES* in italics. Specific table names have been replaced with generic descriptive names to maintain data security. These descriptive names are often correspond with similar tables and fields included in a standard Banner installation that may exist but are generally not used at UVU. Field names have also been changed where the name in the table is sufficiently obscure to make understanding difficult for the reader.

showing data as it exists currently for all terms (including any transactions that affect data for a term after the term has ended, such as retroactive withdrawals from courses). A set of “freeze tables” contain data snapshots allowing time-series analysis throughout a term, and include freezes for the official census and end-of-term reporting dates.

These frozen data from the official reporting dates is used principally for state and federal government reporting. But there is a strong expectation that data reported by the institution for non-government purposes, including that used to make and justify decisions, will be consistent with the government reporting data. For example, between 2010 and 2012, UVU created a web-based data dashboard to provide more specific information on retention and graduation rates than was reported to IPEDS. It nonetheless relied on IPEDS definitions of retention and graduation rates, demographic categories, and reporting cohorts. The cohort definition is especially important, as the IPEDS cohort includes only first-time, full-time, degree-seeking undergraduates entering in fall, who made up only 32% of new UVU students in the 2012-2013 academic year. Because of the expectation that locally-used data will be consistent with government reporting data, the translation regime in place at UVU is defined disproportionately by the rules that govern the three customized government reporting tables.

While certainly the experience with this system is less systematic as a data collection technique than would be ideal, given that one objective of this paper is to establish translation regimes as a theoretical concept for understanding data as a type of social artifact that influences the achievement of social justice, it does not seem unreasonable to interpret that experience using frames and techniques common to emergent methods in social science. The approach used here shares especially some (but not all) features with constructivist grounded theory (Charmaz 2008). This approach is especially appropriate for the study of information systems on three grounds that are especially relevant to the study of information justice:

> [F]irst, it was useful for areas where no previous theory existed; second, it incorporated the complexities of the organizational context into the understanding of the phenomena; and third, that [grounded theory method] was uniquely fitted to studying process and change. (Urquhart 2007, 341)

There are clear parallels between grounded theory and the work presented here. I use an abductive, iterated approach to building theory from empirical research in which both methods of inquiry and substantive findings are emergent rather than predetermined as I move from experience to data structures to interviews, testing the codes and concepts developed previously for consistency with further iterations of inquiry. My approach also, to an extent, works at a distance from existing literature on other problems in information systems and technological ethics in order to avoid artificially constraining the emergence of a broader theory of information justice. (Urquhart 2007, 350–351) However, I must stress that understanding the creation of data using grounded theory was not intent at the outset of this project; grounded theory is itself emergent in this research. It does not rely on the formal data collection processes of open coding and memo writing,[4] nor does it remotely approach theoretical saturation. The latter, of course, provides the opportunity to view this quite weak implementation of grounded

[4] Kelle (2005) and Charmaz (2008) provide exceptional reviews of these specific techniques, defending respectively the two distinct approaches created by the schism between Glaser and Strauss, the founders of grounded theory. While it is impossible to fully argue the point here, I would suggest that the focus on specific methods in that schism has missed the real strength of grounded theory: its reliance on abductively created theoretical concepts that are iteratively tested and refined. This would, especially, be more consistent with the approach's Peircean roots, in which the origin of theory is a creative act and science consists not in the body of knowledge but in subjecting claims abstracted from experience to the examination of further experience. (J. A. Johnson 2000) Specific approaches to code development are not necessary for the success of grounded theory in the same way that, for example, successfully passing tests of statistical significance is for quantitative research using a hypothetical-deductive methodology. From this perspective the test of good grounded theory is its tendency to approach theoretical saturation rather than its compliance with any particular research procedure, and specific coding processes are evaluated from a purely instrumental perspective (i.e., is it helpful for moving toward theoretical saturation). The lack of compliance with such procedures in this paper would thus argue for its inefficiency but not its inadequacy as a work of grounded theory.

theory methodology as an initial iteration of the process and thus valuable as a preliminary approach to the question of what data represents.

3 LEGIBLITY AND THE TRANSLATION REGIME

The ubiquity of data in contemporary society hides its peculiarity. Data is a very specific form of information, one in which the subject is broken down atomistically, measured precisely (in the sense of being measured to quite specific standards that may or may not involve a high level of quantitative precision), and represented consistently so that it can be compared to and aggregated with other cases. That this form of knowledge is more common in highly structured institutions and rose to ubiquity with the modern, bureaucratic state and the capitalist enterprise should surprise no one. Creating data should be regarded as a social process in which reality is made legible (Scott 1998) to the authorities of an institutional structure. Legible knowledge transforms reality into standardized, aggregated, static facts that are capable of consistent documentation and limited to the matters in which there is official interest. Such facts emerge from a process in which common representations are created into which cases are classified and which can then be aggregated to create new facts on which the state will rely in making decisions. (Scott 1998, 80–81)

The need for legibility defines not only the form but also the substantive nature of data. The common scientific realist perspective that sees data as a technical artifact representing objective information about some subject is a quite problematic view of data. It suggests that the process of representing reality[5] is an automatic, even algorithmic process. Such a view is naïve,

[5] For the purpose of this paper, I take “reality” to mean the physically existing world as interpreted by actors within it. Here I follow Charles Sanders Peirce in his seminal essay “The Fixation of Belief” in which there is an underlying reality that cannot itself be perceived but that can be asymptotically approached through repeated observation. (Peirce 1992) This leaves open an

however; like virtually all technologies (J. Johnson 2006), data is a socio-technical construct in which human agency and social structure is central (Nissenbaum 2010, 4–6) and the path from reality to data is contingent rather than determined. (Seaver 2014) Rather than being an automatic process with a one-to-one relationship between reality and data, data states are underdetermined with a one-to-many relationship between reality and data: one state of the world can give rise to many possible data states, some of which are incommensurable with others. In order to make the world legible to human authorities and algorithmic bureaucracies, one data state must be chosen to represent a state of the world from among many possibilities. Reality constrains those possibilities but it does not, by itself, fully reduce the state of the world to a single data state. The contingency of the final forms of data requires some external source of stability in order for data to bring legibility to the world. (Mitev 2005) What is needed is a process of translation from reality to data that constructs a single representation by serving as the external source of stability for representation.

Classifying individuals within a system of gender relations is a good paradigmatic case that demonstrates the operation of a constructivist understanding of data beyond the domain of cultural production. Simply within the binary gender system common in western cultures, people might be represented within a data system either by sex or by gender. These categories are not reducible to each other; the existence of transgendered and intersexed people is sufficient to make sex and gender incommensurable within such binary systems. Moreover, there is no inherent reason that a data system needs to be limited to a gender binary, even in predominantly Western contexts: Facebook recently introduction of more than 50 custom gender descriptions from which its members can choose. (Facebook 2014) The intellectual construct “gender” is thus insufficient to determine how data systems will represent a specific

interpretive space but does not deprive the concept of reality of meaning, allowing it to be bracketed as a distinct but related problem from that addressed in this paper.

person; the reduction of gender realities to specific categories cannot be an objective, value-free, observational process. In spite of this, most data systems rely on the same binary coding frame, one in which gender is taken to have a one-to-one correspondence with biological sex. The representation of individuals' place in the system of gender relations is thus determined by neither reality nor by the technical requirements of the data system. It is a choice on the part of developers to reduce an exceptionally complex reality to a specific legible form.

In order for the process of selecting a data state from among the many possible ones to be, in fact, legible, the process must be a rule-governed one. Creating data from reality is not simply an interpretation but a translation (or, more precisely, a series of translations) in which substantive content embedded in a set of technical rules determines how reality will be represented in the data system. For a relational database,[6] those rules are largely, but not entirely, contained within the data system itself, expressed as technical specifications within the database. The construction of data in relational databases consists mainly in the design and selection of rules such that they implement the demands of the content sources and only secondarily, when the rules and content sources are insufficiently precise, in the direct interpretation of reality by those entering data into the data system. Collectively, one might refer to these structures as the translation regime for a data system.

Translation regimes can be broken down into their technical structures and their social content sources. The technical structures include storage, validation, relation, and extraction structures within the database itself. The most basic storage structure of the relational database

[6] In a more general theory of data, the choice of database type would itself be understood as part of the translation regime. Raman (2012) shows that the choice of a relational database rather than one based on Unstructured Information Management Architecture (UIMA) to maintain land claims in India prevented the storage of knowledge held in narrative form, as was common among *Dalits* in the region. Narrative knowledge would have to be translated into an atomic structure in order to be stored in a relational database; in this case such knowledge was simply excluded in favor of that contained in state-produced documents that could be stored in a relational system. Since all of the data currently used by UVU is contained in relational databases, the influence of database type must be investigated in another context.

translation regime is the structure of individual data tables. The fields selected for inclusion in an individual table do much more than selecting which aspects of reality will be stored (though they most certainly do that as well). Those fields break down that reality into component parts. This is, of course, only a selection of the parts of the reality, and recombining these parts creates only an interpretation of reality rather than an objective and complete representation of it. A validation table contains a list of values that are acceptable for use in a field, used commonly in fields that contain categorical data with a limited number of possible values.

The relationships among data in different tables further shape the translation regime. In a relational database, data tables are structured so that tables can be joined to each other on common elements to allow cases in one table to be matched to cases in another. In the absence of appropriate commons field on which to join, however, data in different tables cannot be related to each other. Data from relational databases is joined and extracted through queries that specify precisely what data will be extracted, how it will be combined in new fields, and how it will be aggregated. A query will, at the least, specify which fields to retrieve for a record, and will usually specify which records to retrieve as well. Applications, whether software systems or analytical processes that connect to a data system, can further translate the data extracted.

While the substantive content of the translations is inscribed in, and to an extent constrained by, the technical structure of the database, the bulk of the substance comes from sources external to the database itself. Culture, the polity,[7] the institution, and private sector actors all provide content for the translations that is then built into the technical structures. As much as the language of conscious design and engineering permeates both the theory and

[7] The multiplicity of meanings of “state” introduces some potential for confusion in this section. In understanding the sources of the specific translation regime studied in this case with the aim of making the concept of translation regimes generalizable beyond the US, it is necessary to refer to both the peripheral political entities of the United States federal system and to the polity generally. In this section, therefore, I will refer specifically to Utah when discussing the former, and use the term “polity” in place of the more common “state” to refer to the latter.

practice of information systems, their conformity with their origin communities' cultural structures suggests that sociological institutions are at least as important. Like many organizational forms, data systems include "not just formal rules, procedures, or norms, but the symbol systems, cognitive scripts, and moral templates that provide the 'frames of meaning' guiding human action" as a mutually-constituting element of social action (Hall and Taylor 1996): data systems are both composed of and instantiate cultural institutions.

Political influences operate in a much more clearly conscious fashion, usually being deliberately designed into the data structures. The polity shapes translations primarily by establishing formal data standards. Data standards define substantively and sometimes technically the content of a data field or record. The translations created by data standards can be quite complex, especially when multiple data standards can apply to the same set of data. Political systems also have more subtle means at their disposal to influence the translation regime as well. Especially for public institutions but, given the public mission of higher education generally, to some extent for all higher education institutions there exists a principal-agent relationship between the polity and those institutions similar to that between legislatures and bureaucracies. That relationship subjects the translation regime to many of the same oversight pressures as any regulatory regime, such as bureaucratic anticipation: agencies, seeing signals from legislators about their desired outcomes, anticipate direction from the legislature and move to secure those outcomes without waiting for that direction to be made explicit (which, in many cases, never happens because the need for direction has been met). (Weingast and Moran 1983) This is not simply having the foresight to see a new formal requirement coming and implement it in advance; it is an act of anticipating the demands of political actors and meeting them as a means of satisfying those actors whether or not the demands are formalized.

The private sector, both for-profit and non-profit, is an important source of content as well. Because UVU's data system is a customized version of a widely used commercial higher education data system, much of the translation regime's content comes from Ellucian, the

makers of Banner, based on a notional higher education institution whose needs are representative of most institutions around which the out-of-the-box version of Banner can be designed. Nor should the non-profit sector's contributions to the content domain of higher education translation regimes should be discounted. Institutions, in a bid to increase transparency (or at least the appearance thereof) are frequently participants in voluntary data sharing processes, each of which comes with their own data standards that may or may not be coordinated with others.

Despite these many external pressures, institutions themselves are important influences on the content of the translation regime. Data standards do not always offer precise operational definitions and logics to determine the data value; they often couple conceptual definitions with a set of valid end states, leaving institutions considerable leeway in the translation process itself. Institutions nearly always control the technical implementation of data standards. Under different alternatives, a particular state of the world can be translated into different values within a data standard depending on how the translation is performed. The institution is also the data collection point, giving it the power to choose both what data to collect and what interactions to translate into data. This is a powerful tool in shaping data: interactions and characteristics that are not turned in to data are not simply missing; they are untranslatable and hence illegible. This prevents them from being considered in decisions.

4 CHARACTERISTIC TRANSLATIONS WITHIN THE REGIME

The data translation regime is not substantively neutral; it favors certain types of outcomes over others. In a relational database such as that used at UVU one can identify at least three characteristic types of translations in the data (as well as, of course, numerous translations that are relatively unique and not analyzed here). These characteristics translations describe how the ontological character and meaning of states of the world commonly change over the course

of the translation process. The result is that translations are most often analytically incommensurable with the reality they purport to express: the words attached to the conditions may be similar, but they are embedded in an entirely new structure.

4.1 Normalizing Translations

One type of translation establishes certain states of the world as part of the realm of normally existing conditions, thus implicitly establishing all other states as deviations from normalcy in some sense. Such translations typically have the effect of reducing the states of the world to only those within the realm of the normal data states. Those represented in the database are thus represented only to the extent that they are capable of being represented within that normal realm; to the extent that they deviate from the normal world as it exists within the database they cease to exist analytically.

The simplest normalizing translation is from relevance to existence. Data is collected based on what the collectors find relevant to their interests: it may shed light on a question they need answered or a decision they may make, or it may be needed to comply with requirements of an external authority. Data is not collected, however, on matters that are not of interest to the institution, nor on matters for which the existence of data is counter to the institution's interests. One common objection to data collection and analysis within IRI was that UVU could be forced to make the data or subsequent analyses of it public under Utah's open records laws. Most frequently this objection was used with projects that might collect data that subjects might consider sensitive but that was not protected by privacy laws, a not unreasonable protection but nonetheless one that is driven by a specific interest on the part of the university. Those characteristics or states of the world were considered irrelevant to decisions, and thus not collected. But when questions arise about such characteristics, irrelevance turns into nonexistence. The characteristics about which there is no data frequently function not as unknowns which need to be estimated or otherwise accounted for in analysis, but are rather

ignored, treated analytically as if they do not exist or, at best, subsumed into platitudes about "context" that fade into the background when the data is available. This is more than just saying that nonexistent data does not exist: it is not data about a given characteristic that is translated into nonexistence but the characteristic itself.

UVU's treatment of religion is an exemplary case. The standard Banner package includes a field for students' religious preferences. UVU does not collect that data from its students, however. Ostensibly this is because of a concern that asking students to identify religious preferences would create the impression that UVU was supporting the dominant religion of its community, The Church of Jesus Christ of Latter Day Saints. This has not prevented UVU's Institutional Research & Information office from including that question on its student opinion surveys, the most recent of which that asked the religion question found that 77.2% of students identify with some form of the LDS faith. (Institutional Research & Information 2013, 45) That data is not included in Banner, however; more than 75% of students' data records have a null value for *RELIGION*. As a result the institution does not routinely consider religion in its decisionmaking, even though such a large number of students sharing a common worldview presents many of the classic problems of in-group/out-group dynamics.

Religion is, to UVU, illegible. This is not at all to say that the institution is hostile to either LDS Church members or non-members; its President, Matthew S. Holland, is the son of one of the highest authorities in the LDS Church and an active church leader in his own right and yet has consistently promoted religious inclusivity toward those outside the LDS Church as an important element of UVU's Core Themes. But the decision not to collect data with which to populate *RELIGION* does leave religious preferences opaque to the institution. The institution cannot ask questions about the role of religion, either as a belief system or as a social institution, in the operation of educational programs. It cannot consider whether students who are not LDS Church members have lower retention rates, a possible sign that they feel excluded from the social life of the campus. It cannot consider whether LDS Church members are less

likely to complete the FAFSA and thus to receive Pell Grants, a possible consequence of a strong ethos of self-sufficiency and financial conservatism within LDS theology and culture. UVU is quite effective addressing these questions within the limits of survey research methods, but a full canvas of students over time is impossible. This leaves UVU unable to "read" a characteristic that student interviews showed is central to many students. Three of the eight students, including all that addressed issues of relatively permanent group membership when discussing their identities, made clear that religion (or lack thereof), always expressed in relation to membership in the LDS Church, is an important aspect of their identities that shapes how they understand their experiences at UVU. Having determined that religion is irrelevant to decision-making and not collected information about it, UVU's students cease, analytically, to have religious preferences.

A similar process takes place with regard to the conditions that a characteristic might take on. Translation regimes transform the diversity of possible conditions of a characteristic into a set of acceptable data values. Those conditions that cannot be represented by a valid data state become represented not as themselves but as deviance: the data is missing; it is given a residual category value such as "other," "not applicable," or "not available"; it is forced into one of the valid data states even if that does not actually represent the state of the world. The validation table for *GENDER* includes only the values "Male," "Female," and "Unspecified," imposing a binary gender schema on the people represented in the field. The "Unspecified" value as a residual is an especially strong reinforcement of the gender binary in this common validation frame: if one is not either male or female (whether because the translation regime insists on correspondence between sex and gender thus denying the existence of transgender identities, or because the person identifies as some form of non-binary identity) one is not even a residual "Other." One is presumed to, in reality, identify with one of the binary values and simply did not communicate that identification to those collecting data.

The USHE data standard for *GENDER* became a stricter one with the inactivation of the "Unspecified" value in the USHE standards in 2012. (Utah System of Higher Education 2013, S–13) This prohibited missing data in *GENDER*. As a result, gender nonconformity is no longer even translated as missing data; all students are translated into one of the binary gender categories. So the diversity of gender identifications are translated into categories of normalcy that are represented by the values "Male" and "Female," and invalid data that exists in a state of deviance from normalcy, first as "Unknown" and then, with the deprecation of that value in the USHE data standards, into a forced choice of a valid but untrue data state. Transgender identities are not simply statistically rare; they are abnormal. And as in the case of irrelevant characteristics of the world, deviant conditions of the world become analytically nonexistent, assumed to be trivial exceptions to a meaningful interpretation of reality.

The translation of irrelevance to nonexistence played a significant role in the creation of UVU's "15 to Finish" program, which encouraged students to take 15 credits per semester in order to graduate in four years. The assumption behind the program is that students who attend full-time are not only more likely to graduate on time; they are more likely to graduate at all. One of the core messages is of the program that it is better to reduce or eliminate outside work in order to attend full-time, even if that requires students to take out loans, because they will be more likely to finish, finish faster (especially within the limits of Pell Grant eligibility), and spend more years earning an income commensurate with their completed degrees. The analysis performed in support of the program did indeed show that this was the case. But it did not consider whether this was practical for all students. UVU does not collect effective data regarding the family status or family income of its students; the only systematic effort at data collection regarding the number of children students have or parents' income that is integrated into Banner is the FAFSA, but institutional privacy protections limit the transfer of FAFSA data outside of the financial aid office and low rates of FAFSA completion make such data unrepresentative in any case. Students with high family incomes might find it much easier to

attend school without working, while those with families might find it especially difficult to reduce or eliminate outside work. Yet neither group exists at an analytical level. The program does have a strong ethos that 15 credit hours may not be appropriate for all students because of their family status or availability of parental support. But that is not implemented formally in the way that, for instance, the various triggers are built into the Stoplight program. “15 to Finish” is for all students, “with exceptions, of course.” These characteristics are irrelevant to the institution’s data collection efforts, become illegible because they are not collected, and ultimately cease to exist as part of the “normal” world in which administrators operate.

It is important especially to understand what it means to say that states of the world *analytically* cease to exist. The qualifier is an important limitation. No one at UVU would deny that many students are religious; the lack of data does not preclude thinking about the characteristic. In a culture where decisions are legitimated in part based on the ability to support them with data analysis, a characteristic that is not datized cannot be analyzed, and so decisions about it cannot be legitimated and are unlikely to be built into policy. Nor can assumptions about the characteristic be questioned. This is perhaps the most pernicious aspect of the translation of relevance. While a characteristic for which data is unavailable may not exist analytically, it may be very prominent culturally, in many cases functioning as part of an ideal type representation and assumed to be true of all cases. The culture of the region carries with it a strong religious identity; survey data showing that a sample of students will be in the neighborhood of 80% LDS confirms this expectation. The result is the assumption that any one student is a member of the LDS Church until they are known to be otherwise.

It is also important to recognize that analytical normalcy is different from social normalcy, by which I mean the existence of certain conditions as the normal or typical condition from which other conditions vary. Self/other distinctions are a form of social normalcy: whites or men represent the normal or typical, while people of color or women are an “other” defined in relation to the norm. The analytical normalcy that I posit here includes both the typical and other

categories in normalcy; deviance constitutes existence outside of the set of recognized categories rather that existence within one of the atypical categories. Analytical normalcy does not imply social normalcy. “White” is, analytically, merely one category of *PRIMARY_ETHNICITY*, not different from other values within the translation regime despite being the only socially normal value. Nor, however, does analytical normalcy challenge social normalcy: the equation of “Male” with normal takes place outside of the translation regime, so that when the translation regime categorizes someone as male or female it does nothing to prevent the substitution of typical and atypical.

4.2 Atomizing Translations

One of the generally accepted best practices of relational database design is that data fields should be atomic, representing one and only one value for one and only one characteristic. To the extent that this is practiced (and it usually is), the result is that translation regimes will represent the world in atomistic terms, fragmenting characteristics that are defined as much by their relationship to other characteristics as by their specific conditions into distinct fields that are not connected to each other. These fields are then analyzed in isolation from each other rather than in the contexts that make them meaningful to the people represented in the database.

Individual identity is highly susceptible to atomization. Complex, intersectional identities frequently bring together different categories of identity into a coherent whole that does not exist within a database: “Jewyoricans” are fragmented into atomistic categories of religion, residence, and ethnicity without the relationships among them that are central to the identity of Jewish New York residents of Puerto Rican descent. These categories reflect both the principle of atomicity—separate fields for separate characteristics—and the data standards to which the institution must conform. The USHE reporting standards for STUDENT maintain separate fields for ethnicity and state of origin (and, of course, does not collect information about religion). (Utah System of Higher Education 2013) This makes representing complex identities that reflect

not just one or another aspect of one's identity but the intersection of or relationships among multiple aspects of one's identity quite rare; data is often analyzed along ethnicity or gender, often sequentially but rarely both at once. There are people represented in UVU's data who are Black, and people who are female; there are some who are both. But there are no Black women in the data.

Atomizing translations can be especially complex when trying to translate a narrative into data. In such cases it is necessary not only to separate characteristics but also to reduce complex conditions into nominal data states that conform to the validation rules and data standards. One might consider the case of students who transfer in large numbers of credits that reflect their prior educational experiences that are at best tangentially related to their current educational ambitions but don't meet the requirements of their current degree program. All of these characteristics are included in STUDENT: *PREVIOUS_EDUCATION* captures whether the student was enrolled at another university in the past, *TRANSFER_CREDITS* reflects the number of hours brought in, *TOTAL_CREDITS* identifies the number of credits earned at all institutions, and *STUDENT_CLASSIFICATION* performs a secondary translation that characterizes overall progress toward the degree. But the fields don't reflect the narrative of a student entering, leaving, and returning with different educational goals and having far more credits than are actually needed to graduate while not being anywhere near completing the current program. A student may be classified as a "Senior" but have perhaps two or three years of additional coursework to complete in order to graduate. The narrative that provides meaning to the value in *TOTAL_CREDITS* is lost; it is reduced to a name: 142.

In most cases, atomizing translations are driven by the content domain rather than the technical domain; the latter merely implements atomicities that are already practiced in other contexts. Technical limitations do not force atomicity on those using extracted data. The different characteristics can be quite easily brought together through simple concatenations of fields or crosstabulations of extracted data. The IPEDS data standards in fact do exactly this.

Institutions are required to report enrollment by ethnicity separately for men and women, (National Center for Education Statistics 2014) allowing the federal government to see the intersectionalty of the two conditions. UVU's Student Success and Retention Dashboard allows analysis by two characteristics at once, making it possible to see the effects of a wide range of two-dimensional intersectionalities, and with some rather awkward technical gymnastics a very narrow set of three-dimensional ones, on graduation and retention rates.

The multi-character *ETHNICITY* field in the USHE data standards and the current IPEDS data standards show how a secondary translation of atomic data can capture the complexity of multi-racial identities. The USHE standard for *ETHNICITY* defines an eight-character field in which each character position represents an ethnicity with which a student might identify, with multiple identifications allowed, chosen among Hispanic or Latino, Asian, Black or African American, American Indian or Alaska Native, Native Hawaiian or Pacific Islander, White, Non Resident [sic] Alien, or Unspecified. (Utah System of Higher Education 2013c, S–14) IPEDS currently used the Office of Management and Budget standards, in which students select all groups with which they identify among American Indians or Alaska Natives, Asians, Blacks or African Americans, Native Hawaiians or Other Pacific Islanders, or Whites and then identify whether or not they are Hispanic or Latino. (National Center for Education Statistics 2014, R) UVU also supports older IPEDS standards that define students by a single ethnicity. To do this, STUDENT includes one binary field for each possible ethnicity that it might report, a count of the total ethnicities selected by the student, and a primary ethnicity to be used with standards that do not support multiple ethnic identities. Those are combined to create both the *USHE_ETHNICITY* and *IPEDS_ETHNICITY* fields. Most on-campus analysis, however, relies on that atomized *PRIMARY_ETHNICITY* field, reinforcing the translation of complex ethnical and racial identities into atomistic categories.

Narratives, too, can be stored in data systems. Banner includes a data table in which comments can be stored. These can provide the narratives that are stripped away by atomizing

translations—if users actually use them. Extracting data from comments is notoriously difficult, requiring complex expressions, tedious analysis, and careful interpretation to make them legible to the institution. Suggesting that the best source for a particular data point is found in COMMENT is universally reviled with UVU's IRI office, but it can be, and sometimes is, done. But this is rarely the case, and even when it is the narrative structure of the comment is rarely used fully, the preference being to identify a nominal value that may be extracted from a comment rather than a more structured field. To be sure, many of these cases involve a re-translation of atomic data. But they do show that other possibilities exist and thus make clear the nature of translation as a choice rather than an inherent technical limitation.

4.3 Unifying Translations

In spite of the imperative toward atomicity, there is a counter-tendency toward unity in translation regimes as well. The detailed and diverse conditions of reality frequently exceed the capability of data systems to store them or the ability of analysts to manage them. A characteristic that can have thousands of potential values, especially when those values are expressed in a nominal level of measurement, does little to bring legibility to the state of the world. Diverse states of the world must often be translated into a small number of values that bring many different conditions together into a common data state.

One translation process that unifies disparate conditions is grouping a large number of possible conditions into a small number of data values. This creates a unified group that may not, in fact, exist in reality or that is at least far more complex than is expressed in a single value label. The USHE ethnicity categories are an example. The standard defines "Asian" as "A person having origins in any of the original peoples of the Far East, Southeast Asian, or the Indian subcontinent including for example Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam." (Utah System of Higher Education 2013, S–14) The common label "Asian" hides a wide range of diversity within the definition;

UVU had Asian students admitted from 27 countries other than the United States among its Fall 2013 students. It seems reasonable to expect that they would have considerable differences among them, and in many cases might find more in common with other racial groups. The borders Pakistan shares with Iran, Afghanistan, and Tajikistan defines Pakistanis as Asians and thus unifies them with students from Japan or Indonesia while separating them from citizens of the surrounding countries who are defined as “White,” a category that includes those “having origins in any of the original peoples of Europe, the Middle East, or North Africa.” Similar differences in racial identity exist between African-Americans and Black immigrants from Africa or the Caribbean and among immigrant groups themselves, (Benson 2006) who are nonetheless unified into a single category of “Black or African American.”

Data systems may also unify characteristics temporally. Characteristics that vary over time become essential and fixed in data systems, stripping away the contingency that is often at work in them. Again, the USHE ethnicity standards are instructive here. The USHE standards make reference consistently to the origins of the student, suggesting that ethnic identity is a fixed part of a person’s overall identity. As a result it can be stored in the data systems and reported consistently over the course of a student’s academic career. But there is considerable evidence that ethnic identity is not essential; rather it is a characteristic that is situated in particular circumstances and can change with them, such as when the student moves from a public to a private space or into and out of spaces dominated by heritage identities. (Zhang and Noels 2013) One might expect this to be especially strong among students who identify with multiple ethnic or racial groups. This situational variability is not captured by the data system, however; the permanence of the data state implies a permanence to the state of the world it purports to represent that may be accurate on average but may not be so at any given moment.

5 CONCLUSION: THE CASE FOR INFORMATION JUSTICE

These characteristic transformations are political acts. The actors that design translation regimes are building structures that embed values and relationships within them that can advantage certain groups over others as the data rather than the actors it represents comes to play a defining role in decision processes. The translation regime begins by representing some groups and excluding other groups, representing some characteristics of individuals but not other characteristics of those same individuals, and representing the data subjects as the data system's designers would represent them rather than as the subjects would. In UVU's data systems, non-credit students and non-degree seeking students do not exist under most circumstances; nearly all queries are designed to filter such students out unless information about them is needed specifically. English as a Second Language students were until recently treated as non-degree seeking and thus left unrepresented in most data-driven decisions. Students' ethnicity is represented but their religion, the most commonly discussed aspect of identity in the student interviews, is not. White students are represented as an ethnicity rather than seeing themselves as ordinary people (who seem to themselves to lack ethnicity), as one White student described himself. These translations are not necessarily hostile to the students' representation, but they do quite clearly shape it.

Just as there are many characteristic translations, there are many political acts that take place through them. The creation of data systems is an assertion of self-interest on the part of the designers; the data system embeds their interests in the decision process but not those who have no influence on the design processes; the latter have no way to make themselves and their interests legible even to institutions that might want to take them into account in good faith, let alone those who might deliberately seek to exclude them. The categorization of characteristics creates and fragments groups that could assert their aims to the institution: Black women are forced to choose to work within the defined fields of *GENDER* and *ETHNICITY* to

meet their needs and thus to accept racial inequality within the feminist movement or gender inequality within Black culture rather than identifying as Black women specifically and pursuing an intersectional strategy. (Hill Collins 2009) Defining states of the world as valid or invalid (e.g., transgender identities) is at the least an imposition of a normalizing judgment through a means other than surveillance, one that has the same kind of potential to construct individuals and groups as hate speech. (Butler 1997) Much of the politics that one would typically expect as groups compete is present in the translation regime.

The politics of the translation regime is different, however, in that it is hidden behind a facade of technical specifications. The translations are, superficially, not exercises of power but simply functions and validation tables that store ostensibly objective information about reality. The scientistic ontology and ideology (Peterson 2003; Haack 1993) embedded in information systems creates the appearance of an apolitical process that is not open to contestation. It thus becomes quite difficult to engage from a political perspective. It cannot be challenged technically, as the translation regime is internally valid and self-legitimating. Any test against reality will confirm the validity of the regime so long as the rules are complied with, because the rules include what data can be considered. Data from within the regime will be correct, and there is no such thing as “data” from outside of the regime. The translation regime creates data; all else is anecdote and thus illegitimate. Challenges to the politics of the translation regime must, then, overcome the issue of legitimacy before the regime can be questioned.

The translation regime is thus a significant and problematic form of political power. Integrating both the technical and the social to render its subjects legible to the exercise of power, the characteristic translations that it produces also exercise power in their own right. As such, the fact that data is constructed through translation, among other processes, presents the need for a theory of information justice. Such a theory must rely on neither controlling the possession of information nor its use. If information is simply representational these would be adequate safeguards. Privacy rights could protect transfer of information, and substantive

regimes similar to human subjects protections might prevent against harmful uses. But the constructive nature of data makes these inadequate. Neither privacy use ethics addresses the content of information that is, within the internal framework of the translation regime, accurate. These approaches cannot address the questions that arise in building data systems in ways that their translations further rather than undermine the individuals represented in them. Instead, a theory of information justice should be oriented toward understanding data as a socio-technical system, promoting design practices that minimize their potential for domination and oppression.

REFERENCES

Benson, Janel E. 2006. “Exploring the Racial Identities of Black Immigrants in the United States.” *Sociological Forum* 21 (2): 219–47. doi:10.1007/s11206-006-9013-7.

Butler, Judith. 1997. *Excitable Speech: A Politics of the Performative*. New York: Routledge.

Charmaz, Kathy. 2008. “Grounded Theory as an Emergent Method.” In *Handbook of Emergent Methods*, edited by Sharlene Nagy Hesse-Biber and Patricia Leavy, 155–70. New York: The Guilford Press.

Facebook. 2014. “How Do I Edit Basic Info on My Timeline and Choose Who Can See It?” Accessed February 25. https://www.facebook.com/help/276177272409629.

Haack, Susan. 1993. *Evidence and Inquiry: Towards Reconstruction in Epistemology*. Oxford, UK: Blackwell.

Hall, Peter A., and Rosemary C. R. Taylor. 1996. “Political Science and the Three New Institutionalisms.” *Political Studies* 44 (5): 936–57.

Hill Collins, Patricia. 2009. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. 2nd ed. Routledge Classics. New York: Routledge.

Institutional Research & Information. 2013. “Student Omnibus Survey -- Fall 2012 Results.” http://www.uvu.edu/iri/documents/surveys_and_studies/Omnibus%20Student%20Survey%20-%20Fall%202012%20Results.pdf.

Johnson, JA. 2006. “Technology and Pragmatism: From Value Neutrality to Value Criticality.” In *Western Political Science Association Annual Meeting*. Albuquerque, New Mexico. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2154654.

Johnson, Jeffrey A. 2000. “Abductive Inference and the Problem of Explanation in Social Science.” In Chicago.

Kelle, Udo. 2005. "'Emergence' vs. 'Forcing' of Empirical Data? A Crucial Problem of 'Grounded Theory' Reconsidered." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 6 (2). http://www.qualitative-research.net/index.php/fqs/article/view/467/1000.

Mitev, Nathalie N. 2005. "Are Social Constructivist Approaches Critical? The Case of IS Failure." In *Handbook of Critical Information Systems Research: Theory and Application*, 70–103. Elgar Original Reference. Northampton, Mass: E. Elgar Pub.

National Center for Education Statistics. 2014. "The Integrated Postsecondary Education Data System - Glossary." Accessed March 10. http://nces.ed.gov/ipeds/glossary/.

Nissenbaum, Helen. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Staford, California: Stanford Law Books.

Peirce, Charles S. 1992. *The Essential Peirce: Selected Philosophical Writings*. Edited by Nathan Houser, Christian J. W. Kloesel, and Peirce Edition Project. Bloomington: Indiana University Press.

Peterson, Gregory R. 2003. "Demarcation and the Scientistic Fallacy." *Zygon* 38 (4): 751–61. doi:10.1111/j.1467-9744.2003.00536.x.

Raman, Bhuvaneswari. 2012. "The Rhetoric of Transparency and Its Reality: Transparent Territories, Opaque Power and Empowerment." *The Journal of Community Informatics* 8 (2). http://ci-journal.net/index.php/ciej/article/view/866/909.

Scott, James C. 1998. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.

Seaver, Nick. 2014. "On Reverse Engineering: Looking for the Cultural Work of Engineers." *Medium.com*. January 30. https://medium.com/anthropology-and-algorithms/d9f5bae87812.

Urquhart, Cathy. 2007. "The Evolving Nature of Grounded Theory Method: The Case of the Information Systems Discipline." In *The SAGE Handbook of Grounded Theory*, edited by Antony Bryant and Kathy Charmaz, 339–60. Los Angeles; London: SAGE.

Utah System of Higher Education. 2013. "Student Data Submission File, 2013-2014 Submission Year." http://higheredutah.org/wp-content/uploads/2013/09/rd_2013DataDict_Students.pdf.

Weingast, Barry R., and Mark J. Moran. 1983. "Bureaucratic Discretion or Congressional Control? Regulatory Policymaking by the Federal Trade Commission." *Journal of Political Economy* 91 (5): 765–800.

Zhang, R., and K. A. Noels. 2013. "When Ethnic Identities Vary: Cross-Situation and Within-Situation Variation, Authenticity, and Well-Being." *Journal of Cross-Cultural Psychology* 44 (4): 552–73. doi:10.1177/0022022112463604.