DATA QUALITY: THE BASICS

February 28, 2017 14:14

INTRODUCTION TO DATA QUALITY

Having a faculty activity reporting solution is important, but it’s even more important to ensure that the data which resides within the database is rich and accurate, so that your users have confidence in the reports generated—from the institution level, all the way down to a specific individual. In fact, achieving a high level of data quality is arguably the biggest key to success with Activity Insight. It unlocks its value, enabling you to trust your data to make informed decisions whenever questions about faculty productivity arise.

Because data quality is so important, we’ve conducted a wide array of research in this area so that we could provide one-on-one assistance to you and to develop overarching strategies to tackle issues at the root of data quality problems. Through this research, we have determined that there are four dimensions which contribute to “good data quality”. These dimensions are represented below:

To determine whether or not a client, or a unit within a client, has quality data, we must measure each of these dimensions individually:

- **Completeness** is the measure of whether data needed to feed your desired outputs are indeed present.

- **Accuracy** ensures that the data entered are truthful and mistake-free. This is the only bucket that Digital Measures is unable to measure as part of our data quality scores. It is up to each client to find an appropriate measure for accuracy, and to audit the data entered at an interval which ensures that the data are consistently clean.
- **Consistency** in data collection is what enables you to trust that all activities of a certain type are entered the same way, in the same place, and that they can be extracted consistently to feed any report that requires that information. It also enables you to count on there being one source for all information of a certain type.

- **Currency** is a measure of the degree to which your data is up-to-date and reflects your faculty’s most recent activities. The more you can count on Activity Insight to provide “fresh” information, the more often it will be seen as the go-to tool for one-off and just-in-time answers to questions about activities and accomplishments.

**MEASURING DATA QUALITY**

Our study looked at data quality for the last one and five years across three of the above dimensions - completeness, consistency and currency. We considered differences in the content and importance of various activities across academic disciplines and factored those into our analysis.

For each dimension we established one or more key metrics, a target for each metric, and a weight. A target is a result that indicates data quality, based on the results for successful clients. The weight represents the relative impact of the metric on overall data quality, for the given timeframe.

**DATA QUALITY DIMENSION: COMPLETENESS**

To measure completeness, we established three metrics:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Target</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median records per user per year for “core screens”</td>
<td>9 records</td>
<td>40%</td>
</tr>
<tr>
<td>Percent of “core fields” populated in records for “core screens”</td>
<td>98%</td>
<td>30%</td>
</tr>
<tr>
<td>Percent of dated records for “core screens”</td>
<td>97%</td>
<td>10%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>Target</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median records per user per year for “core screens”</td>
<td>X records</td>
<td>36%</td>
</tr>
<tr>
<td>Percent of “core fields” populated in records for “core screens”</td>
<td>98%</td>
<td>27%</td>
</tr>
</tbody>
</table>
Percent of dated records for “core screens”

<table>
<thead>
<tr>
<th></th>
<th>97%</th>
<th>9%</th>
</tr>
</thead>
</table>

* X: This target is determined based on the types of unit(s) that make up your instrument

**Metric 1: Median records per user per year for “core screens”**

This metric sheds light on the degree to which the records in Activity Insight represent your faculty’s activities by measuring the presence of records for the most essential, or core, screens. Core screens are those that provide the bulk of the content for annual and accreditation reports, among many other potential outputs of the system.

**Selected Core Screens:**

- General Information
  - Yearly Data
  - Awards and Honors
  - Consulting
  - Education

- Teaching
  - Directed Student Learning
  - Scheduled Teaching

- Scholarship/Research
  - Artistic and Professional Performances and Exhibits
  - Contracts, Fellowships, Grants and Sponsored Research
  - Intellectual Contributions
  - Presentations

- Service
  - Service to the Institution
  - Professional and Public Service

**Metric 2: Percent of “core fields” populated in records for “core screens”**

This indicates the likelihood that records will display in a meaningful way in reports run from Activity Insight. Meaningful representation of activities in reports requires that the records contain data in the most essential, or core, fields. To be conservative the bar was set low with core fields; you might consider many more fields to be essential. Our list, however, is one that we believe to be non-negotiable - a list of fields that, without which, the information is of questionable use.
Metric 3: Percent of dated records for "core screens"

Dates are critical for ensuring that records are correctly included in reports. The percentage of dated records for core screens illuminates the volume of records that can be accurately included in reports run for a specific timeframe.

DATA QUALITY DIMENSION: CONSISTENCY

As a starting point for studying this dimension of data quality, we focused on one behavior that generates data consistency - entering data as a series of discrete, specific elements instead of in a narrative or block. In this model, each granular piece of data about an activity record - like first name, middle name, last name - has a specific home. Within the realm of granular data collection, specific field types - like drop-down lists, text boxes, numeric fields, etc. - further encourage consistency. This is where we focused our attention, investigating how often two specific field types are used as intended.

To measure consistency, we established the following metric:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Target</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of records for &quot;core screens&quot; that use &quot;Type&quot; and &quot;Description&quot; fields as intended</td>
<td>95%</td>
<td>20%</td>
</tr>
</tbody>
</table>
As indicated, we studied the use of two types of fields:

- "Type" fields, which are predefined drop-down lists used for categorization
- "Description", "Comments" and "Notes" fields, which are text area fields used to capture facts that have no other, more specific, home

For Type fields, we looked at two ways they are commonly used and misused.

- For records with a Type field available, how many actually have that field populated?
  - **Effective use**: A user selects a Type for her Contracts, Grants, Fellowships and Sponsored Research record.
  - **Ineffective use**: A user adds a Presentation record to the system but while doing so does not select a Presentation Type from the drop-down list.

- For records with both a Type and a related Explanation of "Other" field, how often are those fields effectively used together?
  - **Effective use**: A user goes to Activity Insight to enter a record for a dataset created as an output of research, but on the Intellectual Contributions screen finds that "Dataset" is not an option in the Contribution Type drop-down list. She selects the "Other" value in the list, and enters "Dataset" into the Explanation of "Other" field.
  - **Ineffective use**: A user goes to Activity Insight to enter a record for a journal article. On the Intellectual Contributions screen, she selects "Journal Article" from the Contribution Type drop-down list and enters "Academic Journal Article" into the Explanation of "Other" field.

For Description, Comments, and Notes fields we looked at how often a record with a value in one of these text area fields also has all of its core fields populated. If all core fields are populated, this indicates that the user embraced granular data entry. This implies effective use of text area fields. If, however, one or more core fields are not populated, the record does not pass the test; it might be evidence that the Description, Comments or Notes field contains summary data, like a citation, that should be parsed into other fields on the screen. The latter is a case of ineffective use of Description, Comments or Notes fields.

**DATA QUALITY DIMENSION: CURRENCY**

To measure currency, we established the following metric:
In an effort to understand how fresh the data in Activity Insight are and the degree to which users embrace the goal of continuous - “as activities occur” - data entry, we studied the volume of records per user with dates that fall within subsets of the larger data quality timeframes.

For the one-year data quality timeframe, “current” records are those with dates within the last four months. A good result and high score for this timeframe means that your users are inclined toward keeping the data up-to-date as activities occur. The better the score and result, the better prepared your data are to provide real-time answers to questions about faculty productivity.

- Regarding the weight for this metric for the one-year data quality timeframe - The currency score for the one-year data quality timeframe has no weight toward the one-year overall score. This is because we have come to understand through our research that continuous data entry is still aspirational for most of our clients; annual data entry in preparation for annual reports is far more common. Giving a weight to the currency metric for the one-year timeframe would have skewed our understanding of data quality and reduced the meaningfulness of the assessment overall by setting an unfair expectation based on future, not current, reporting objectives.

For the five-year data quality timeframe, “current” records are those with dates within the last one year. The lower the result relative to the target, the more data entry must take place between now and when you need to run reports on the last year’s data.

WRAP UP

Data quality scores are reviewed and updated on a quarterly basis. Contact your Success Consultant to retrieve your most up-to-date packet, and continue to do so periodically to ensure that you monitor the health of the data which exists in your system.

After you’ve analyzed your scores, review the following articles in the “Data Quality” section of the Resource Center which provide best practice advice on improving your
scores, based on the data quality dimension(s) where improvement is necessary:

- Improving Your Data Quality | Completeness
- Improving Your Data Quality | Consistency
- Improving Your Data Quality | Currency

In addition, be sure to review our articles on the fourth, unmeasured dimension of data quality: Accuracy.

- How to Measures Data Quality Accuracy
- Improving Your Data Quality | Accuracy

After having reviewed these articles, work with your Success Consultant to determine best strategies for your individual institution.