Agenda

- Data Governance Website
- Business Glossary
- Data Definitions for Student Records
- Data Accuracy/Validation
Agenda

- **Data Governance Website**
- Business Glossary
- Data Definitions for Student Records
- Data Accuracy/Validation
The Data Governance Initiative has its own website, and it can be found at www.nyu.edu/ir/datagov.
This is what the website looks like.
On the upper left side we have a listing of NYU resources.
On the lower left side are copies of the slides from the Data Steward Advisory Committee meetings.
The top right provides other NYU web resources.
And the bottom right shows external resources dealing with data governance and data quality.
Data Governance Website

www.nyu.edu/ir/datagov
Agenda

• Data Governance Website
• **Business Glossary**
• Data Definitions for Student Records
• Data Accuracy/Validation
A Business Glossary was the top request from the Data Stewards when we polled you at the first meeting.

In order to take advantage of existing University resources, we are going to build the Business Glossary using the Knowledge Base portion of ServiceLink. This will be similar to the DSG/PSO efforts to develop “How To” articles for ServiceLink.

We anticipate the process will go something like this:
- use workflow within ServiceLink to move definitions along
- work with Data Stewards to develop definitions of data elements
- the definitions would be entered into ServiceLink
- the completed definitions would be forwarded to an “Editorial Board” for review/editing
- after polishing, the entry would be published to ServiceLink

We are still looking into ways the Glossary might be accessed. One thought, by no means certain, would be to highlight a term to be looked up, or perhaps have some simple URL to look up term.

The expectation is that we will move existing definitions, developed by the PSO/DSG, into the new NYU Glossary. We will likely work with a few departments at the outset to “shake out” the new system. We will be in touch with Data Domain Trustees to finalize plans.

We hope to have a prototype for the next Data Steward meeting. In the meantime, here is what we envision NYU Glossary might look like.
This is the outline of the Business Glossary entry.
There would be an additional page, with information of a more technical nature. It would not automatically display so as not to be overwhelming, but it would be available for those who might need the information.
This is an example of what the page might look like from within ServiceLink.
And this is the rest of the general page.
This is what the top half of the technical page might look like from within ServiceLink.
And this is the bottom half of the technical page.
Here’s another example. Note that there is a hotlink out to another website. This will allow the Business Glossary to take advantage of other resources both within and outside of the University.
Examples
0401.00 (Biology, General)
0502.00 (Accounting)
0502.10 (Taxation)

Synonyms

See Also
CIP Code

Data Steward
Office of the University Registrar

Click here for technical information on this field
Technical Glossary: HEGIS Code

Data Classification
Public

Source System
PeopleSoft Campus Solutions
New York State Education Department, Office of Research and Information Systems

Source Description
HEGIS codes are assigned by New York State when a new academic program is approved for the inventory of registered programs.

NYU Institutional Research And Data Integrity
On-Line
PeopleSoft Campus Solutions
Set Up SACR
Foundation Tables
Reporting Codes
HEGIS Code Table

Database Location

<table>
<thead>
<tr>
<th>Environment</th>
<th>Table</th>
<th>Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEBPROD</td>
<td>SISCAMPUS HEGIS_CODE_TBL_V</td>
<td>Hegis Code</td>
</tr>
<tr>
<td>EPMPROD</td>
<td>PS_NYU_HEGIS_TAXONOMY</td>
<td>NYU HEGIS_CODE</td>
</tr>
</tbody>
</table>

UDW+ Reporting

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Folder</th>
<th>Data Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIS - Term Registration</td>
<td>CPP/Organization HEGIS Code</td>
<td></td>
</tr>
</tbody>
</table>

Click here for general information about this field
As Student Records is next area moving into the UDW+, we began working with the University Registrar to develop definitions for those data fields which are being incorporated into their new dashboards and ad hoc reporting files.

The Registrar felt that it wouldn’t be possible to define data fields simply based on their names, as many fields have similar sounding names. In addition, the Registrar is familiar with data fields on the screens within the on-line system, and the names used on the screens, and not necessarily with the more technical data field names in isolation.

The Student Record Data Definition Working Group thought about developing a lineage trail in a pretty manual method. The NYU IT representative on the working group suggested that NYU IT might be able to develop an automated lineage. They are probably sorry they said that now, as it is possible, but it is a large effort. While developing a Proof of Concept for the tool, NYU IT realized that such a lineage would be helpful to them as well.
Data Definitions for Student Records

- OBIEE (Subject Area/Folder/Data Field)
- Warehouse Reporting Table
- OWS Layer (Staging)
- Campus Solutions Tables
- Campus Solutions Screens

There are two endpoints: the OBIEE data field within a folder within a subject area AND the data field as seen and labelled on the screen.

There are five steps being documented.

With the data lineage identified, when working on developing definitions, one could start at either end and trace how the data moves through the various systems, including any calculations or transformations. This will be a great aid in developing definitions, understanding how each OBIEE field is derived, and where the data comes from.

NYU IT is looking at this effort as a pilot, to understand the efforts involved and engagement needed from all parties – particularly NYU IT and business data stewards. We can then build a longer term strategy to build this type of a tool for other areas.

As we had seen on the Business Glossary slides, we are making provision for this information to be included in the NYU Business Glossary.
Data accuracy and validation was ninth on the list of top 10 items you identified as potential areas for improvements in dealing with data. One way of identifying “bad” data is by conducting Data Profiling.

And there are lots of examples of bad data in our systems:
There are potential duplicate records. It would seem that it is possible we have two cases of two vendor entries for the same vendor.

Is this a problem?
- If one of the two entries for a vendor were updated, would a payment get to the proper place?
- A summary of payments to a vendor wouldn’t aggregate appropriately
- Could report more vendors than actually exist
And then there are “fake” records.

Minnie Mouse is probably not a real student.

What are the problems here?

- If Minnie is in the system as an applicant, there could be bad counts
- If Minnie is in the system and registered in a course, there could be bad counts and a needed seat might not be available to a real student.
And then there is incomplete or out-of-bounds errors.

This is from a data profile of Student Term Data. This has identified 10 records (out of 3.8 million records) which probably need to be reviewed since it is difficult to imagine how a student might enroll for less than 0 points in classes.

What is the potential impact?
- an incorrect average point count
- is there bad data elsewhere? Number of registrants?

Bottom line: bad data is everywhere in our systems, sometimes not too obvious, but always potentially throwing off some information which one of your managers may be asking for.

Let’s spend a few minutes thinking about data accuracy.
Like to open it up for discussion. Here are two questions for you. Looking for detailed answers

Data Accuracy/Validation

What are the data accuracy or validation problems you face in your office – either from data you generate, or from data you receive from other offices?
And – what type of business rules might catch that bad data?
Thank You!