DOI: 10.36315/2019v2end073
Education and New Developments 2019

TOWARDS HIGHER EDUCATION DATA HYGIENE – A CASE STUDY

Ji Hu, & Xu Chu Meng
Chancellors' Office, New York University Shanghai (China)

Abstract

While developments in data analytics have provided unprecedented opportunities for Higher Education Institutions (HEIs) to understand themselves and fulfill their missions, it remains a challenge to maintain good institutional data hygiene, with well-known barriers including lack of incentives beyond regulatory compliance, and thus lack of data sharing within the institution (Krawitz, Law, and Litman, 2018). The paper is a case study of the efforts of NYU Shanghai to improve institutional data hygiene when building the infrastructure for analytical transformation, with its nature as a Sino-US joint venture adding to the complexity. By addressing the above-mentioned challenges through establishing a central workgroup, actively utilizing regulatory reports, and engaging operational units into the process of defining and visualizing institutional data structure, NYU Shanghai made significant progress, including successfully initiating its Data Warehousing Project to integrate data silos and to improve institutional data hygiene at large. The paper also discusses how the experience of NYU Shanghai may be applied to other HEIs at a similar stage of data maturity.

Keywords: *Higher education institutions, data hygiene, regulatory compliance, data silos, analytical transformation.*

1. Background

Developments in data analytics have provided unprecedented opportunities for Higher Education Institutions (HEIs) to understand themselves and fulfill their missions. However, often underrepresented in the discussion is the importance of building a foundation for analytical transformation through institutional data hygiene (Jones, 2017), the collective processes to guarantee data integrity (Rouse, 2013), to which well-known barriers include lack of incentives beyond regulatory compliance, and thus lack of data sharing within the institution (Krawitz, Law, and Litman, 2018).

The objective of this paper is to first, through the case study of the efforts of NYU Shanghai towards institutional data hygiene, provide an example in this regard for other HEIs at a similar stage of data maturity. It also aims to raise awareness, and initiate a comprehensive discussion, both in and out of the university, on the risks of poor data hygiene.

2. NYU Shanghai and its data characteristics

New York University Shanghai (NYU Shanghai) was jointly founded in 2012, by New York University in the United States and East China Normal University in China. Located in Shanghai, the university enrolled its inaugural undergraduate cohort in 2013, and graduated its first class in 2017. It has so far grown into a campus with about 1300 full-time undergraduate students and 200 full-time faculty.

The institutional data at NYU Shanghai were characterized by three underlying factors. Being a comprehensive research university indicates a wide scope of data domain areas and a complex data structure, while the small scale allows more accuracy and resilience in locating and addressing institutional data hygiene crux. Besides, as the first Sino-US joint venture, NYU Shanghai has a unique dual identify, being both an independent university in China, and part of the global network of NYU as a degree-granting campus, which not only requires compliance with both systems, but also poses additional challenges to the dynamic data processes.

The university initiated its efforts towards institutional data hygiene in 2017, as the first step of its attempt to build the infrastructure for analytical transformation. The efforts were facilitated at this time for two reasons. After going through an entire student life cycle from application to graduation, NYU Shanghai had obtained the knowledge and experience of a fully-functional university, as well as the

institutional data generated in the cycle, and had equipped itself for such data initiatives. Besides, it was also before the data issues accumulated at the bottom of operational systems, providing more flexibility in addressing the identified issues.

The data characteristics at NYU Shanghai make it a suitable subject for this case study, taken into consideration the generalizability of its experience to other HEIs due to its comprehensive and complex data landscape; the feasibility of in-depth analysis of institutional data hygiene from the small scale and lack of historical burden; and also the specificity of the discussions in which it is involved, derived from its dual identity, in the wave of the internationalization of higher education.

3. Towards institutional data hygiene

The initial efforts of NYU Shanghai to improve institutional data hygiene were unfolded in three parts, respectively establishing a central workgroup, actively utilizing regulatory compliance reports, and engaging operational units into the process of defining and visualizing institutional data structure. The three elements intertwined with each other to address the institutional data hygiene challenges by incentivizing and enabling data sharing and integration among data silos.

3.1. Establish a central workgroup

Similar to Conway's law that "organizations which design systems … are constrained to produce designs which are copies of the communication structures of these organizations", data silos are usually created by operational silos within the institution, where lack of interactive collaboration beyond minimum sufficiency is preventing active data sharing and integration.

In order to connect the data silos, an Institutional Research Workgroup (IRWG) was established at NYU Shanghai in 2017, led by university's Associate Vice Chancellor. With a wide range of centralized data initiatives on its agenda, the workgroup has also been serving as a liaison among operational units, both administratively and statistically. The core of the connection is a reciprocal process, where IRWG addressed data integrity issues directly through integrating data from different operational units and performing data audits, and indirectly through collecting data needs from operational units, and eliminating needs for quality-threatening shadow systems, which are siloed spreadsheets or databases, for example, stored on user's computer, usually derived from the failure of central systems to meet their needs.

Instead of instilling the function in an existing unit, the new IRWG is functionality-driven, enabling it to push data flow among units beyond traditional practice, and to integrate unit-specific needs into a larger central initiative.

3.2. Actively utilize regulatory reports

Instead of articulating a new data mandate to add to the busy daily routine of operational units, NYU Shanghai adopted a different strategy to actively utilize regulatory compliance reports that had already been on the task list of the units.

The establishment of IRWG provided an opportunity to systematically investigate the multiple regulatory reports beyond the capacity of any single operational units. A series of steps targeting institutional data hygiene were employed, with the above-mentioned data audits being the initial step. After the audit of the regulatory reported data, there were two major findings in addition to successfully identifying data integrity issues. The first was that the regulatory reported data highly overlapped with core institutional data, thus making it a good representation of institutional data status and a start point for data hygiene efforts. The second was a reinforcement of data silos in the reporting process, where operational units independently provided data in their own subject areas, which had resulted in struggles in fulfilling the requests, especially from both Chinese and US sides that did not naturally reconcile with each other. A re-visit of the reporting process, with IRWG coordination at the center, formed a larger picture for all operational units to better know how they fit into the puzzle of institutional data, instead of being isolated in their own.

Being an existing and struggling task for operational units, the active utilization of regulatory reports not only incentivized data sharing among units, but also brought the units to work together.

3.3. Engage operational units in visualizing institutional data structure

Despite the essence of IRWG in forming centrality of communication, the successful engagement of operational units in the efforts towards institutional data hygiene, from the beginning, was key to the accomplishments of the university.

The subject matter expertise of operational units was leveraged in the process of investigating institutional data status, and visualizing institutional data structure. IRWG conducted several rounds of communication with operational units for these purposes, and multiple units in different stages of the same data pipeline were gathered to discuss the matter together, such as Academic Affairs and Registrar on student study-away and course election. The communication guaranteed that the conclusions are based on coherent and shared understandings.

As the frontline data custodians, operational units also had more ownership and accountability in maintaining data hygiene through their direct engagement in visualizing institutional data structure.

4. Accomplishments and implications

Through the above-mentioned efforts, NYU Shanghai has made significant progress towards better understanding of its data, and thus better institutional data hygiene. An Institutional Core Data Inventory has been developed, which serves as a uniform and sole source of reference for core institutional data items, a visualization of the institutional data structure, and a documentation of the metadata. Besides, Institutional Data Reporting has been coordinated and streamlined to both guarantee data integrity and address business priorities. The efforts culminated in the initiation of the first institution-wide Data Warehousing Project, which would provide the infrastructure to reinforce the accomplishments so far by avoid system-perpetuating data silos, and also lays the foundation for further analytical transformation of the university.

The experience of NYU Shanghai can be applied to other HEIs at a similar stage of data maturity, in following aspects. Both technical and subject matter expertise need to be leveraged and brought together in these efforts, as they respectively hold perspectives of methodology enhancement and business priorities. An interconnecting part between the two expertise is extremely beneficial in terms of communication efficiency and standardization, as well as targeting immediate business needs to gain buy-in from operational units. Last but not least, advanced data users in operational units are valuable resources and pioneers that can drive changes in their own units if engaged appropriately.

Acknowledgements

We would like to thank NYU Shanghai operational units for their contributions to the institutional efforts in improving data hygiene, which is core to this case study. We also extend our special gratitude to the other members of IRWG for their guidance and support. We appreciate the joint efforts with IT to initiate and implement the Data Warehousing Project. The study benefited from author's attendance at 2019 Higher Education Data Warehousing Forum (HEDW), which was jointly funded by NYU Shanghai and HEDW Cathy Lester Attendance Grant.

References

Jackson, N. (2018). Connecting data silos in higher ed. *University Business, 21(4)*, p. 40. Retrieved February 12, 2019, from https://universitybusiness.com/connecting-data-silos-in-higher-ed/.

Jones, H. (2017). Data hygiene part 1: how dirty data hurts your bottom line. Retrieved April 1, 2019, from https://www.linkedin.com/pulse/data-hygiene-part-2-how-can-make-your-business-more-profitable-jones/.

Kim, J. (2018). Steps to higher education data cleanliness: a technical perspective. Retrieved April 6, 2019, from https://evisions.com/resources/blog/data-cleaning-technical-perspective/.

Krawitz, M., Law, J., & Litman, S. (2018). How higher-education institutions can transform themselves using advanced analytics. Retrieved March 6, 2019, from https://www.mckinsey.com/industries/social-sector/our-insights/how-higher-education-institutions-can-transform-themselves-using-advanced-analytics.

Rouse, M. (2013). Guide to managing a data quality assurance program. Retrieved March 1, 2019, from https://whatis.techtarget.com/definition/data-hygiene.