เทคนิคการปรับเปลี่ยนและการทำความสะอาดข้อมูล
(กระบวนการพัฒนาระบบสารสนเทศ)
ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สป.

ด้วยศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร สำนักงานปลัดกระทรวงศึกษาธิการ มีบทบาทหน้าที่ตามกฎกระทรวง ในการพัฒนาระบบคลังข้อมูลและฐานข้อมูลสารสนเทศ รวมทั้งเครือข่ายเทคโนโลยีสารสนเทศและเป็นศูนย์กลางข้อมูลของกระทรวง ซึ่งการดำเนินการในการพัฒนาระบบข้อมูลสารสนเทศด้านการศึกษา โดยการประสาน รวบรวม จัดเก็บข้อมูลสารสนเทศจากทุกหน่วยงานทั้งในและนอกสังกัดกระทรวงศึกษาธิการ รวมทั้งสิ้น ๙ กระทรวง ได้แก่ กระทรวงศึกษาธิการ กระทรวงคมนาคม กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์ กระทรวงกลาโหม กระทรวงการท่องเที่ยวและกีฬา กระทรวงวัฒนธรรม กระทรวงมหาดไทย กระทรวงสาธารณสุข และส่วนราชการไม่สังกัดสำนักนายกฯ หรือกระทรวง โดยมีการจัดเก็บข้อมูลพื้นฐานด้านการศึกษา มาตั้งแต่ปีการศึกษา ๒๕๔๖-ปัจจุบัน มีการบริหารจัดการรวมถึงระบบจัดเก็บข้อมูลอย่างเป็นระบบ โดยนำเทคโนโลยีสารสนเทศที่ทันสมัยเข้ามาช่วยในการบริหารจัดการและสามารถให้บริการข้อมูลได้อย่างมีประสิทธิภาพ

ในปีงบประมาณ พ.ศ.๒๕๕๘ ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร ได้ร่วมกันพิจารณาทบทวนการจำแนกรายการองค์ความรู้ของทุกกระบวนการหลัก และคัดเลือกองค์ความรู้ที่สนับสนุนการดำเนินงาน มาจัดทำแผนการจัดการความรู้ โดยองค์ความรู้ที่จำเป็น คือ “เทคนิคการปรับเปลี่ยนและการทำความสะอาดข้อมูล” ของกระบวนการพัฒนาระบบสารสนเทศ เนื่องจากปัจจุบันกระบวนการ รวบรวม และประมวลผลข้อมูลเข้าสู่ระบบฐานข้อมูล ยังขาดบุคลากรที่มีความชำนาญ และบุคลากรส่วนหนึ่งยังขาดทักษะในการจัดการระบบฐานข้อมูล ทำให้การประมวลผลข้อมูลมีความล่าช้า ส่งผลกระทบต่อการให้บริการข้อมูลสารสนเทศด้านการศึกษาของหน่วยงาน จากแผนการจัดการความรู้ (KM Action Plan) คณะทำงานได้ร่วมกัน บ่งชี้ความรู้ สร้างและแสวงหาความรู้ จัดกิจกรรมการแบ่งปันแลกเปลี่ยนความรู้ มีการประมวลผลและกลั่นกรองความรู้ เพื่อหาวิธีการปฏิบัติที่ดีที่สุด ดังนี้

ก่อนดำเนินการ

๑. เลือกวิธีที่เหมาะสมสำหรับการทำความสะอาดข้อมูล ซึ่งวิธีการทำความสะอาดข้อมูล ขึ้นอยู่กับความสะอาดของข้อมูล ว่ามีความสะอาดมากน้อยเพียงใด ซึ่งในกระบวนการ ETL หมายถึง การสกัด (Extraction) การส่งผ่าน/แปลงข้อมูล (Transformation) และการนำข้อมูลเข้า (Loading) วิธีทำความสะอาดข้อมูลที่นิยมใช้ ดังนี้

- ทำความสะอาดเบื้องต้นโดยการตรวจสอบด้วยสายตา แล้วจึงทำความสะอาด
- ทำความสะอาดโดยใช้ Tools เช่น Function VLookUp ของ Microsoft Excel, เขียนโปรแกรม หรือใช้ Regular Expression

๒. พิจารณาขนาดของข้อมูล ซึ่งมีส่วนสำคัญหากเป็นข้อมูลขนาดใหญ่ ไม่เหมาะที่จะแก้ไขด้วยมือ หากเป็นข้อมูลปริมาณน้อย สามารถตรวจสอบด้วยสายตา และแก้ไขหรือจัดรูปแบบให้เป็นไปตามรูปแบบในระบบฐานข้อมูล ได้ทันทีโดยไม่ต้องใช้ Tools

๓. มี Data Dictionary ซึ่งเป็นสิ่งสำคัญที่จะทำให้รู้ว่าข้อมูลแต่ละคอลัมน์ แต่ละฟิลด์ เป็นข้อมูลอะไร มีหน่วยวัดอย่างไร ค่าของข้อมูล ประเภทข้อมูล ซึ่งอาจมีการกำหนดชื่อเขตของข้อมูลแตกต่างกัน เนื่องจากข้อมูลมาจากหลายแหล่ง เช่น รายการข้อมูล “เพศ” บางหน่วยงานใช้ Sex บางหน่วยงานใช้ Gender เป็นต้น

การดำเนินการทำ Data Cleaning โดยเลือกใช้วิธีที่เหมาะสม

๑. สำหรับผู้ปฏิบัติงานทั่วไป มีวิธีการ ดังนี้

๑) ตรวจสอบด้วยสายตา ด้วยการ print out ข้อมูลมาตรวจสอบ ทำค่าที่หายไปปรับให้มีความ Smooth หากมีข้อมูลสูญหายให้แทนค่าข้อมูลด้วยการเติมค่าคงที่ค่าหนึ่ง (ค่าที่กำหนดขึ้น) หรือเติมค่าที่หายด้วยมือเปล่า

๒) การแยกคำ หากได้รับข้อมูลที่มีทั้งคำนำหน้าชื่อ ชื่อ-สกุล อยู่ร่วมกัน ต้องแยกออกออกจากกัน

๓) ตัดเครื่องหมาย หรืออักขระที่ไม่มีความหมายออก เช่น ID มี (-) หรือ เว้นวรรค ต้องตัดเครื่องหมาย หรือลบช่องว่างออก

๔) การแปลงรูปแบบวันที่ วันเดือนปีเกิด ต้องจัดให้มีรูปแบบตามที่กำหนดในฐานข้อมูลคือ ป/ด/ว หากเป็น ค.ศ. ต้องทำเป็น พ.ศ. ระมัดระวังการแสดงค่าใน Excel ที่มักแสดงค่าผิดจากเดิมโดยอัตโนมัติ

๕) ข้อมูลที่อยู่ ต้องแยกคำ บางหน่วยงานมีเลขที่และชื่อตำบล อำเภอ จังหวัด รวมอยู่ในฟิลด์เดียวกันทั้งหมด อาจต้องใช้เวลาในการกรองหลายครั้ง

๖) ข้อมูลตัวเลข ต้องระมัดระวัง หากนำไปใช้งานต่อ เช่น รหัสที่มีเลข “๐” อยู่ข้างหน้า ต้องแปลงรูปแบบตัวเลขให้เป็นตัวอักษรก่อน จึงจะแสดงผลครบถ้วน

๗) กรองข้อมูลหลายครั้ง จนกว่าจะได้ข้อมูลที่ต้องการ ทั้งนี้อาจใช้ Function VLookUp ของ Microsoft Excel ที่ผู้ปฏิบัติงานทั่วไปสามารถใช้งานได้

๘) จดบันทึกความผิดพลาด ในแบบฟอร์มบันทึกข้อมูล

๒. สำหรับโปรแกรมเมอร์ หรือนักวิชาการคอมพิวเตอร์ มีวิธีการ ดังนี้

- ใช้ Text Editor Tools ใช้ในการกรองข้อมูลเบื้องต้น เช่น Edit Plus, Notepad+, Eclipse Visual Studio ทำการ save และ Import เข้า Database

- Regular Expression เป็นเครื่องมือที่ช่วยในการจัดกลุ่ม/ค้นหาข้อความที่ต้องการ เช่น ข้อมูลเป็น “สป.ศธ.” และ “สป ศธ” อ่านเหมือนกันแต่ความหมายต่างกัน ต้องมีเทคนิคในการแยก/กรองข้อมูลโดยกำหนด pattern ของข้อความที่ถูกต้อง และดำเนินการเอาช่องว่าง ตัวอักขระที่ไม่จำเป็น ที่ไม่สื่อความหมาย (.) (,) (') (/) (-) ออกและใช้การ replace คำที่ถูกต้องอีกครั้ง มี Database เพื่อเก็บ log การประมวลผลเพื่อนำกลับมาดูได้อีกครั้ง

๓. จดบันทึกขั้นตอนโดยละเอียด หากมีการทำกระบวนการซ้ำ คนอื่นสามารถทำได้

การดำเนินการหลังการทำ Cleaning Data

๑. นำข้อมูลเข้าสู่ระบบฐานข้อมูลตามรูปแบบที่กำหนด

๒. รวบรวมข้อผิดพลาด หารือร่วมกับหน่วยงานที่เป็นผู้จัดส่งข้อมูล

ข้อควรระวังเพื่อมิให้เกิดข้อมูลที่ไม่สะอาด

๑. กำหนดมาตรฐานรายการข้อมูล และมีการปรับปรุงอย่างสม่ำเสมอ

๒. ต้องมีการตรวจสอบหรือปรับปรุงรายการข้อมูลหลักให้ทันสมัยอยู่เสมอ

๓. สร้างความตระหนักของผู้ปฏิบัติงานเพื่อให้เกิดคุณภาพของข้อมูลที่ดี

“ผู้ทำหน้าที่ CLEANING DATA ต้องมีความละเอียด รอบคอบ