Public vs. private data

back up to How to plan data ingest for VIVO

previous topic: Typical sources of VIVO data

Public vs. private data

Whenever possible, leave non-public information out of the public VIVO, since including private information will complicate a user's picture of his or her VIVO profile and make the entire project more difficult to manage. Semantic web tools have been developed to share data by exposing it for direct consumption in other tools as well as for human eyes to read, and while the Vitro software underlying VIVO offers ways to limit the visibility of the data on websites, a complete RDF export of a VIVO database will be directly readable by other tools that may make no attempt to filter by any criteria.

As with any data and any software, this is a common sense balance of benefits and risks. There are many reasons to include data such as department identifiers in VIVO that should be hidden from view to avoid clutter but are essential for aligning new data; the project will add more ways for users to limit the visibility of certain research-related information a person may not wish to share, such as a network of informal colleagues or a new area of investigation. However, we see little to gain and much to lose by putting any confidential data into VIVO, such as salary history, termination dates, leave status, or identification information (age, sex, race, nationality, marital status, home phone or address, etc). Cornell's VIVO instance links users to the campus directory rather than holding contact information directly, since our HR system frequently lags employees' own updates of their contact information.

Should VIVO become a System of Record (SOR) at your institution? That's really up to you, but you need to carefully consider the risks as well as benefits. VIVO may well become the SOR for information such as research areas and keywords, brief statements of research purpose, and perhaps publications. For other information such as grants and appointments that are currently maintained elsewhere for administrative purposes, VIVO should remain a downstream consumer of SORs rather than seeking to supplant core systems. At Cornell the college administrators feel a pressing need to have a data mart that combines all the information they need about faculty, including HR, grants, courses, course evaluations, and assorted other information including some they track directly. They have wide-ranging requirements for running reports on that data, however, and need to include salary history, grades, performance reviews, and other data that would be much better managed through a data warehousing and report generation tool behind appropriate firewalls than by a VIVO instance designed for public information discovery and sharing.

---

next topic: Ingest tools: home brew or off the shelf?