

How to manage data cleanup in VIVO

General principles of data cleanup as they apply to VIVO

Anyone who's built an enterprise-scale application, whether for a large or small enterprise, sooner or later has to address malformed, incorrect, and just plain stale data. And even though the problems may originate elsewhere, your application surfaces them and hence may have to take some heat.

What to do?

- Acknowledge the problem – and try to explain briefly the nature and source of the problem.
- Correct problems at the source if at all possible – while sometimes you have to make an emergency correction or update, the problem may be reintroduced with your next update if it isn't fixed at the source.
 - Every VIVO should plan for [Emergency Removal of Data \(Example of Policy\)](#) in response to legal or administrative action
- When errors originated in VIVO or will not get overwritten by later automated updates, empower and train people to correct them, either through self editing or proxy editing. Not everyone will, but those that learn how like having that level of control – and if it's an admin aide who's taken the brunt of a complaint from a faculty member, then he/she will be happy to be able to report an instant correction.
- Report bugs to the [Mailing Lists](#).
- Add principles based on your experience here

Issues with manual corrections

- When you need to correct the spelling of an address, or the name of a journal, be aware that you are editing the address or journal itself, and not the person or organization's connection to that shared address – so you will need to follow the link in VIVO to that separate entity to do the editing
- Publications may be assigned to the wrong author, and where two people have the same name it may be difficult to select the correct one from the pick list
- Will this correction be applied to the data source of reference? Does the error return at the next ingest?
 - Some sources are ingested only once per semester, so it is difficult to track whether corrections are overridden.
 - Because the data is not primarily maintained to benefit VIVO, the source office may not be responsive to VIVO's issues.
 - Again, a personal relationship with the data administrator can be very valuable.