AdministrativeStatistics

Administrative Statistics

We talked a bit at DSUG 2005 about administrative statistics being a third kind of report (alongside activity and archive stats). We identified two main areas that could benefit from having some administrative stats associated with them:

- Submission
- Workflow
  and for reasons that will become clear shortly, I would like to add:
- Archived item activity
  Thinking about this further, there are two main realizations:

1. There is no need to have a separate logging system for administrative statistics. These can be inferred from the log files after the event provided that:
   a. The item_id being acted upon is available for every logging action we are interested in
   b. The logging events for the administrative areas are kept consistent. Any change in, for example, logging action name would break any implementation of this type
2. That the OAIS reference model suggests maintaining usage statistics for AIPs, and that this sort of material could provide at least some of that information. Therefore, we would attach per AIP reports on activity throughout the life of the item as an additional AUDIT bundle or similar.

an example of how this might work

If we imagine an item passing through submission and then a full three stage workflow, the activity of the item would be roughly as follows:

3. A blank item is created and a db id assigned
4. the user edits the metadata for the item
5. the user adds bitstreams to the item
6. the user accepts the licence
7. The item enters the workflow
8. WF1 administrators carry out checks on the item (few logging events worth noting)
9. WF2 adminsitrators modify the item, add/remove bitstreams
10. WF3 adminsitrators modify the item, add/remove bitstreams
11. The item enters the archive
    1-4 would be submission and 5-9 workflow. (Note that if the workflow becomes configurable to some higher degree in the future, we may have to think again about which logging actions are recorded, but a sufficiently flexible log file analyser should be able to make sense of this without too much modification (if any at all)).
    Once this full submission process has completed there will be a set of actions associated with an item id in the log files/database, and we would be able to simply trawl the data, reconstruct the sequence of actions on a particular item, and using a little bit of knowledge of how we want to present this information we could build reports on a per-item basis for everything in the archive. Using the list of likely areas that we want to associate administrative statistics with, we would want to record the following activities:

- Submission
  - Start and end times of submission
  - actions carried out (separated by user, since there may be collaborative efforts)
- Workflow
  - Start and end times of WF1, and views carried out by which users
  - Start and end times of WF2, views carried out, bitstreams added, metadata modified by which users
  - Start and end times of WF3, views carried out, bitstreams added, metadata modified by which users
  - Time item becomes available in the archive
- Archived item activity
  - item views (with times)
  - bitstream downloads (with times)
  - information provided by distinct individual IP addresses visiting

what this means for the stats system

Basically it means that we can not worry too much about logging administrative statistics at this stage, since they are reconstructible from the log files. Instead we would want to build specific administrative stats analysis tools for the reporting end of the system instead. There are things to consider such as performance, and whether to run these statistical reports on-the-fly or have them periodically generated. I like very much the idea that we generate flat files which are then attached to the item itself for long term preservation. We could make the files either directly in XHTML or some intermediate XML format (perhaps we could use DocBook here, as per the Documentation Tasks page). This would mean that when specific statistics are requested for an item we wouldn't run a query against the database, but instead simply display the part of the item relevant to the stats.