Rapid Cataloguing of Data: Managing data with GUIDs and File Properties

The following are characteristics of project data for which this method is targeted.

* Same data appearing in multiple files
* Dynamic environment, i.e. new files being created, existing ones renamed or moved to different locations, some files deleted etc.
* Discernible patterns of file and folder naming, e.g. data in folder ASB are all associated with the Alternatives to Slash and Burn Project
* Data shared across different computers

At ICRAF the following are projects with these characteristics.

* GRP6 database
* Partnership database
* Proposals database
* WKIEMP
* Carbon Benefits Project
* Training database

The following constitute the initial steps of rapid cataloguing

1. Harvesting files and their properties to a working database from the project data space.
2. Filtering useful data files from noise, resulting in Unique files (Upath).
3. Assigning GUIDs to data files. The basic principle is that similar data in multiple data files should be assigned the same GUID
4. Writing back files to disk along with their associated GUIDs (immediately)
5. Cataloguing the GUID-assigned data files using
   * Manual methods
   * Path patterns
   * User interface

Rapid cataloguing in for the target data is a repeated operation; so it matters how each of the above steps is implemented initially and in subsequent iterations. The following notes should guide the detailed implementation of specific steps.

Filtering, step 2 is not bullet proof. So discarded files may in fact be valid data; on the other hand, some non-filtered filed may need to be discarded. Users must be able to control this situation

The following scenarios should influence the implementation of step 3

* When you find out that a brand data new file has been added to the project data space
  + Will it be assigned a new GUID or one of the existing ones?
* When you discover that 2 data files with different GUIDs have in fact the same data
  + Merging GUIDs
* When you discover that 2 files with the same GUIDs have different data
  + Splitting GUIDs

We already have encountered the first scenario. I expect the others to arise as well, so you need to plan well about them.