# ILC Data Catalog

A data catalog has been created to provide a navigable repository for ILC files. The database structure allows for arbitrary metadata to be associated with the files, and includes a download link.

## Using the ILC Crawler to Add Files to the Data Catalog

A Java crawler is being developed to automatically add existing STDHEP and SLCIO files to the database, and use the filenames and event headers to collect metadata which can be stored alongside it.

The metadata collected (where available) is:-

- Energy
- Polarization
- SLIC and GÉANT version
- Number of events
- Detector
- A list of all the collections stored within a SLCIO file.

At present, the program simply takes each file supplied to it and generates a shell script of commands using the data catalog API's registerDataset method to add entries to the catalog.

### Usage

To generate the shell script, simply call the Crawler class and supply the location of the file(s) as a parameter from the command line:-

```
java ilccrawler.Crawler <files>
```

This will create a file crawlerscript.sh, which when run will add the files to the database.

#### Example

```
java ilccrawler.Crawler /nfs/slac/g/lcd/ilc_data2/ILC250/LOI_higgs/sid02/slcio/slic/*.slcio
```

This produces a shell script, the first lines of which are as follows:-

```
#!/bin/bash
~srs/datacat/prod/datacat registerDataset --define nEnergy=250 --define "polarization=(+80, -30)" --define
"slicversion=slic-v2r5p3" --define "geantversion=geant4-v9r1p2" --define nEvents=1000 --define "Detector=sid02"
--define "Collections=VtxEndcapHits:TkrBarrHits:MCParticle:HcalEndcapHits:LumiCalHits:TkrForwardHits:
EcalEndcapHits:TkrEndcapHits:MuonEndcapHits:VtxBarrHits:MuonBarrHits:EcalBarrHits:BeamCalHits:HcalBarrHits:
MCParticleEndPointEnergy:" SLCIO ILC/250/slcio/slic /nfs/slac/g/lcd/ilc_data2/ILC250/LOI_higgs/sid02/slcio/slic
/125fb-1_+80e-_-30e+_higgs_run1-000-0-1000_SLIC-v2r5p3_geant4-v9r1p2_LCPhys_sid02.slcio
~srs/datacat/prod/datacat registerDataset --define nEnergy=250 --define "polarization=(+80, -30)" --define
"slicversion=slic-v2r5p3" --define "geantversion=geant4-v9r1p2" --define nEvents=1000 --define "Detector=sid02"
--define "Collections=MCParticleEndPointEnergy:LumiCalHits:MCParticle:TkrBarrHits:MuonEndcapHits:EcalBarrHits:
TkrForwardHits:VtxEndcapHits:TkrEndcapHits:VtxBarrHits:HcalEndcapHits:EcalEndcapHits:MuonBarrHits:BeamCalHits:
HcalBarrHits:" SLCIO ILC/250/slcio/slic /nfs/slac/g/lcd/ilc_data2/ILC250/LOI_higgs/sid02/slcio/slic/125fb-
1_+80e-_-30e+_higgs_run1-000-1-1000_SLIC-v2r5p3_geant4-v9r1p2_LCPhys_sid02.slcio
~srs/datacat/prod/datacat registerDataset --define nEnergy=250 --define "polarization=(+80, -30)" --define
"slicversion=slic-v2r5p3" --define "geantversion=geant4-v9r1p2" --define nEvents=1000 --define "Detector=sid02"
--define "Collections=TkrEndcapHits:VtxBarrHits:MuonBarrHits:MCParticle:MCParticleEndPointEnergy:MuonEndcapHits:
TkrForwardHits:VtxEndcapHits:BeamCalHits:EcalBarrHits:EcalEndcapHits:LumiCalHits:HcalBarrHits:HcalEndcapHits:
TkrBarrHits:" SLCIO ILC/250/slcio/slic /nfs/slac/g/lcd/ilc_data2/ILC250/LOI_higgs/sid02/slcio/slic/125fb-1_+80e-
_-30e+_higgs_run1-000-10-1000_SLIC-v2r5p3_geant4-v9r1p2_LCPhys_sid02.slcio
~srs/datacat/prod/datacat registerDataset --define nEnergy=250 --define "polarization=(+80, -30)" --define
"slicversion=slic-v2r5p3" --define "geantversion=geant4-v9r1p2" --define nEvents=1000 --define "Detector=sid02"
--define "Collections=EcalEndcapHits:LumiCalHits:MCParticle:MCParticleEndPointEnergy:MuonBarrHits:
TkrForwardHits:EcalBarrHits:VtxBarrHits:HcalEndcapHits:VtxEndcapHits:HcalBarrHits:TkrEndcapHits:MuonEndcapHits:
TkrBarrHits:BeamCalHits:" SLCIO ILC/250/slcio/slic /nfs/slac/g/lcd/ilc_data2/ILC250/LOI_higgs/sid02/slcio/slic
/125fb-1_+80e-_-30e+_higgs_run1-000-11-1000_SLIC-v2r5p3_geant4-v9r1p2_LCPhys_sid02.slcio
...
```

As can be seen, each file has its own command, printed on a separate line, with its own metadata (which in this case are the same as the files listed are different runs of the same simulation).

The script can is then run:-

```
crawlerscript.sh
```

Each file is then registered in the catalog. The final result can be seen at http://srs.slac.stanford.edu/DataCatalog/folder.jsp?folder=573739.

### Limitations

The process of parsing each file, and then adding them to the catalog is a lengthy one and for large directories will take several minutes.

There is also a bug which dumps several "Not creating second sensor" error reports to the command line every time the crawler checks a file's event header.