THREDDS Catalog Generator

Ethan Davis

UCAR/Unidata

Problem

Catalog generation is tedious when have more than a handful of datasets

```
<collection name="NCEP AVN-Q" dataType="GRID“
            serverID="motherlode">
  <dataset name="2002-05-03 12:00:00 UTC“
           urlPath="dods/model/2002050312_avn-q.nc" />
  <dataset name="2002-05-03 06:00:00 UTC“
           urlPath="dods/model/2002050306_avn-q.nc" />
  <dataset name="2002-05-03 00:00:00 UTC“
           urlPath="dods/model/2002050300_avn-q.nc" />
  <dataset name="2002-05-02 18:00:00 UTC“
           urlPath="dods/model/2002050218_avn-q.nc" />
   …
</collection
```

Goal

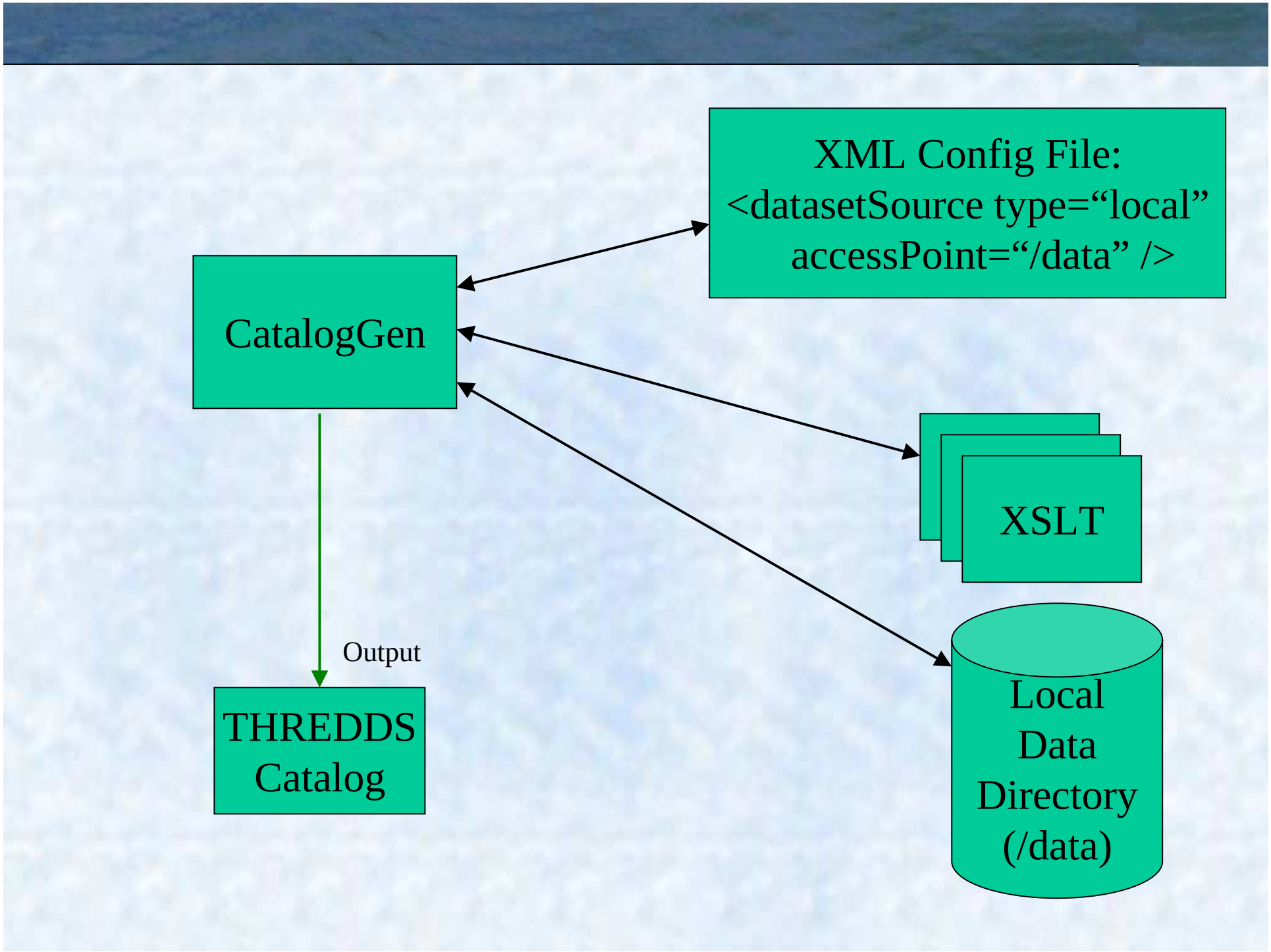
- Automate catalog generation as much as possible

First Generation: picatAG

- Currently generating Unidata model data catalog on motherlode
- Java Application using JDOM on XML configuration file
- Scans local directories
- Hardwired for InvCatalog 0.4
- Hard to maintain

Current Generation: CatalogGen

- Java Application using XSLT on XML configuration file
- Scans local directories
- Flexible: can generate THREDDS Catalogs or Aggregation Server Config file
- Modular, easy maintenance


XML Config File:
<datasetSource type="local"
accessPoint="/data" />
CatalogGen
XSLT
Output
THREDDS
Catalog
Local
Data
Directory
(/data)

Catalog Example: Setup

```
<?xml version="1.0" encoding="UTF-8"?>
<catalog name="Unidata IDD Server - Model Data" version="0.5">
  <serverList />
  <collection name="Model data" dataType="Grid">
    <datasetSource name="first" type="local" structure="flat“
          accessPoint="/htdocs/motherlode/dods/model">
      <resultServer type="DODS" serverID="mlode“
          base="http://motherlode.ucar.edu/cgi-bin/dods/nph-nc/“
          accessPointHeader="/htdocs/motherlode/" />
    </datasetSource>
  </collection>
  <datasetNamer type="RegExp“
      matchPattern=
        "([0-9][0-9][0-9][0-9])([0-9][0-9])([0-9][0-9])([0-9][0-9])_avn-q.nc$“
      substitutePattern="NCEP AVN-Q $1-$2-$3 $4:00:00 GMT" />
</catalog>
```

Catalog Example: Result

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE catalog SYSTEM
"http://www.unidata.ucar.edu/projects/THREDDS/xml/InvCatalog.0.4.dtd">
<catalog name="Unidata IDD Server - Model Data" version="0.4">
   <server type="DODS" ID="mlode“
      base="http://motherlode.ucar.edu/cgi-bin/dods/nph-nc/"/>
   <collection name="Model data" dataType="GRID">
      <dataset name="NCEP AVN-X 2002-04-27 12:00:00 GMT“
         urlPath="dods/model/2002042712_avn-x.nc"/>
      <dataset name="NCEP RUC 2002-05-02 00:00:00 GMT“
         urlPath="dods/model/2002050200_ruc.nc"/>
      <dataset name="NCEP RUC 2002-05-02 12:00:00 GMT“
         urlPath="dods/model/2002050212_ruc.nc"/>
      <dataset name="NCEP RUC 2002-05-03 12:00:00 GMT“
         urlPath="dods/model/2002050312_ruc.nc"/>
       …
   </collection>
</catalog>
```

Agg Config Example: Setup

```
<aggServerConfig name="Example Agg Server Catalog“ version="0.4“>
   <server type="netCDF" ID="local" base="file:/mlode-data/"/>
   <aggregation serverID="local" variable="" type="2">
      <datasetRef xlink:href="#COADS-1deg"/>
         <fileSource type="local" accessPoint="/mlode-data/model">
            <resultServer type="local" serverID="local“
               base="file:/mlode-data/" accessPointHeader="/mlode-data/" />
         </fileSource>
   </aggregation>
   <catalog name="Example DODS Aggregation Server Catalog“
        version="0.4" dataType="GRID" serverID="this">
      <server type="DODS" ID="this“
         base="http://thredds.unidata.ucar.edu:8080/dodsC/"/>
      <collection name="(Type 2) COADS">
         <dataset name="1degree Equatorial“
            urlPath="COADS-1degree-equatorial" id="COADS-1deg"/>
      </collection>
   </catalog>
</aggServerConfig>
```

GDS Example

- GDS 1.2 XML Catalog at COLA (link)
- Run (link)

  “java thredds.cataloggen.CatalogGen –gdsCat http://cola8.iges.org:9191/dods/xml”
- Demo (link)

Issues/Weakness

- Requires human setup
- Limited to scanning local file
- Doesn’t know anything about data (filename manipulation only)
  - Can’t assign dataType
  - Can’t assign Dataset Descriptors

Future Plans (short-term)

- Expand directives language
  - Add <datasetFilter> (grouping datasets)
  - Add <fileNamer> (used for Agg type 1, variable value)
- Clean up
  - Allow sorting of datasets (new directive?)
  - Move defaults up hierarchy (serverID from dataset to collection)
- Improve handling of GDS 1.2 XML catalogs

Future Plans (long-term)

- Crawl DODS servers, DODS file servers
- User interface (human guidance)
  - Build XML input files
  - Additional metadata
- How will Schemas affect things

Possible Discussion Issues

- Other sources to crawl
  - DODS servers, DODS file servers
  - DODS datasets for additional metadata
  - LAS config files (CGI access)
  - Other???
- Automating addition of other metadata
- “Human in the loop” automation (UI)