Data formats in e-Science

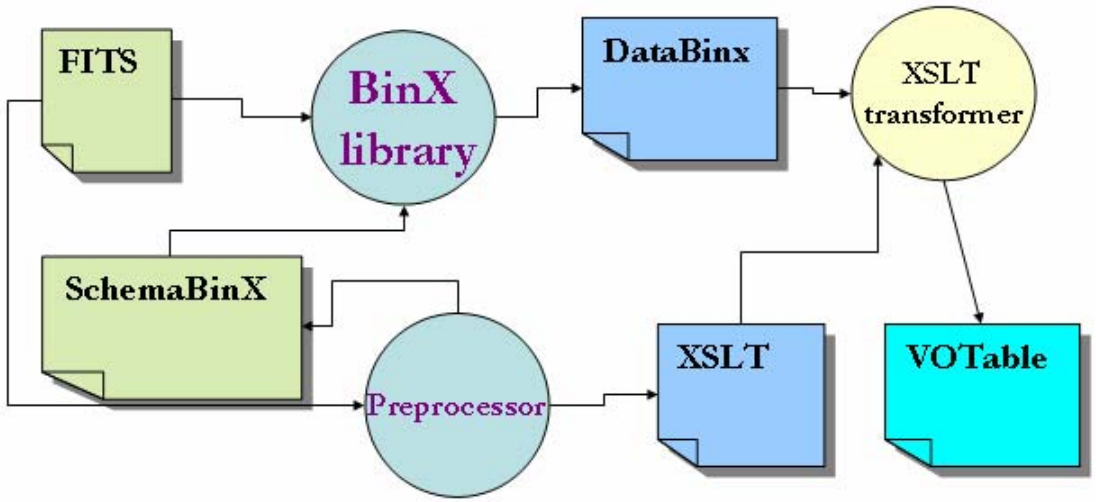
- Two key requirements
  - Interoperability and Scalability
  - **XML** is flexible, but verbose
  - **Binary formats** are compact, but specific
  - e.g. VOTable vs FITS issue in the VO
- Two possible solutions:
  - BinX from the *edikt* project at NeSC
  - VX from the School of Informatics

BinX: Binary in XML

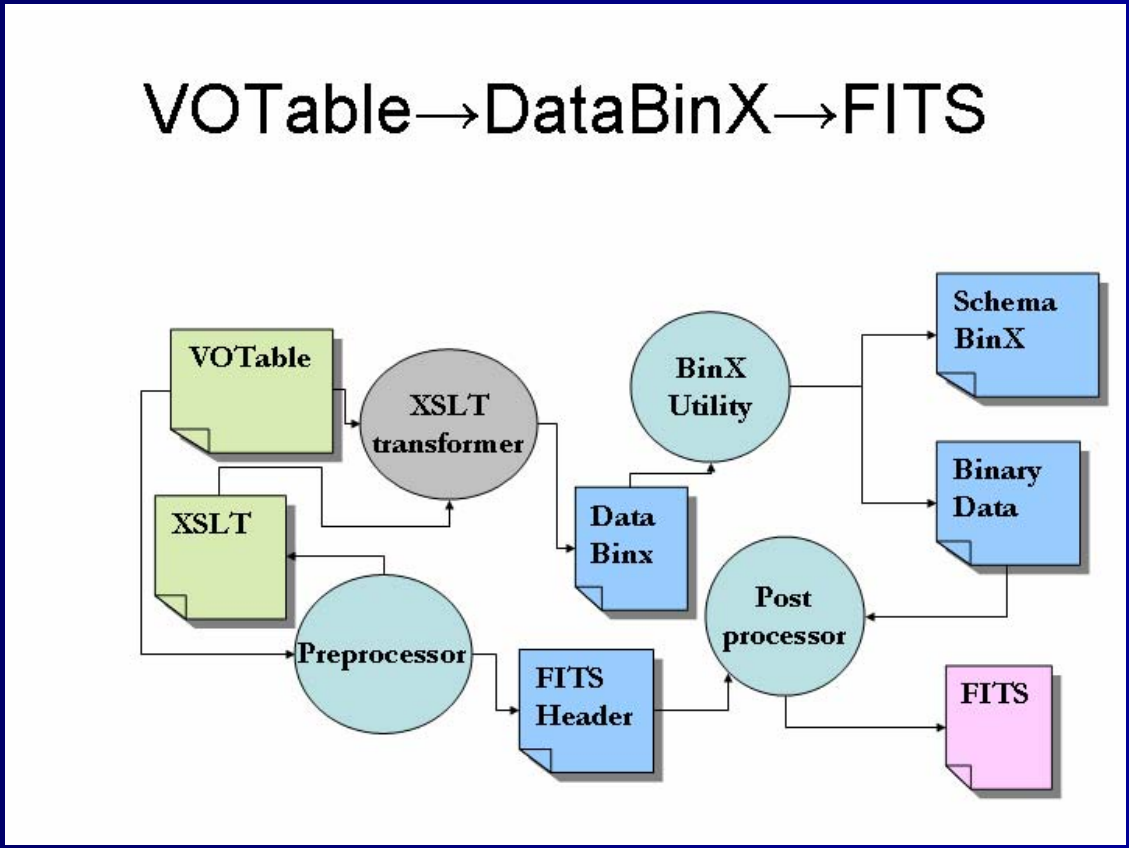
- A language:
  - Uses XML to describe the data types and structures in a binary data file
- A library:
  - For manipulating XML and binary files
- BinX files:
  - SchemaBinX: XML descriptor of binary file
  - DataBinX: SchemaBinX + data values
- BinX will allow you to interact with a binary file as if it were XML – e.g. run XPath queries

Astronomical testbed: format conversion with BinX

FITS →DataBinX →VOTable



...and back again



This way is harder, due to ASCII text in FITS header.

More about BinX

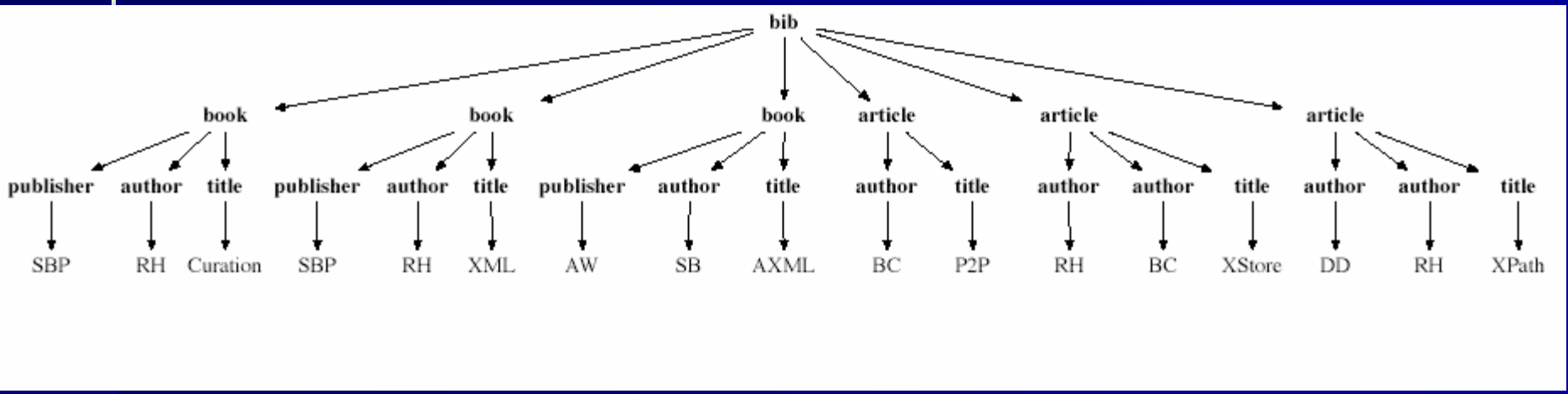
- Download the BinX code & play with it
  - See *www.edikt.org/binx*

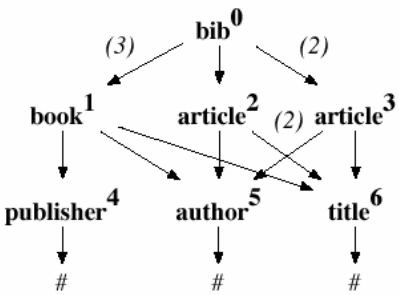
- After BinX comes DFDL (*daffodil*)
  - Data Format Description Language
  - Developing through GGF Working Group
  - BinX might morph into a DFDL implementation
- Basic idea behind DFDL:
  - We can't have a single data format, but we can have a single way of describing data formats

VX: Vectorizing XML

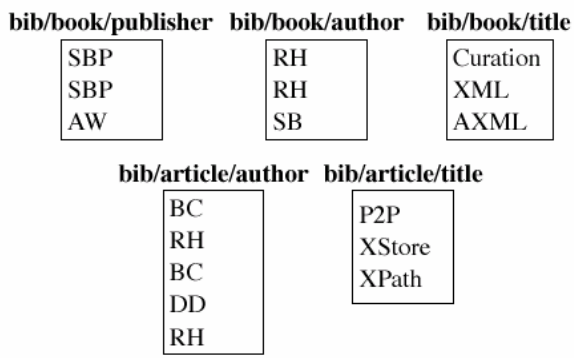
- XML seems especially verbose for data files with simple, repeating structures
  - e.g. VOTable – lots of <TR>s and <TD>s
- Vectorize it:
  - Decompose the XML document into a *skeleton* describing the structure and *vectors* containing data values

Example: bibliography in XML





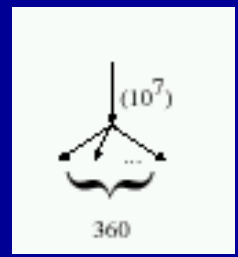
(a) Skeleton S



(b) Data vectors

Astronomical VX application

- Export the PhotoObj Table from the SDSS EDR database into VOTable
  - 360 columns and 10,000,000 rows
- The Skeleton here is trivial



- Querying the decomposed XML version can be as fast as querying the SkyServer database
  - For queries where SkyServer doesn't make heavy use of indexes, which are not in VX yet

More: http://homepages.inf.ed.ac.uk/v1bchoi/paper.pdf