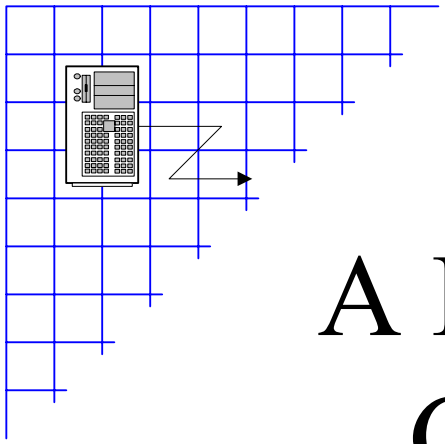
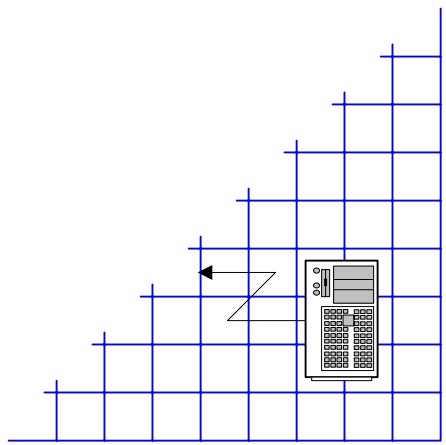


A Hybrid XML-Relational Grid Metadata Catalog

Scott Jensen, Beth Plale, Sangmi Lee Pallickara, and Yiming Sun

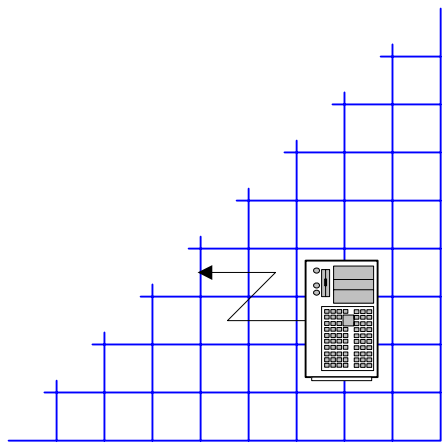
DDE Lab

Indiana University



Metadata and Grids

- Importance of multidisciplinary federated curated collections of data
  - NSF Blue-ribbon panel on cyberinfrastructure
  - Central Laboratory of Research Councils CLRC)
  - FGDC and National Spatial Data Infrastructure (NSDI) clearinghouse.
- Capturing metadata as it is generated.

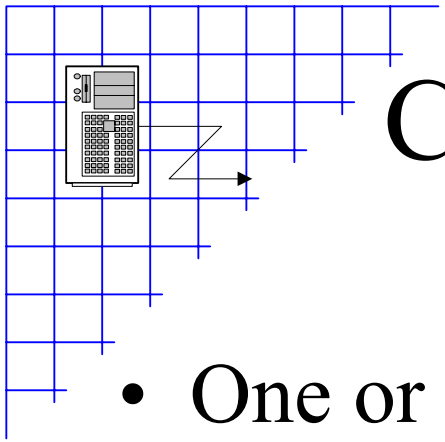


LEAD Overview

- Large Information Technology Research (ITR) project that is creating an integrated, scalable cyberinfrastructure for mesoscale meteorology research and education.
- A major underpinning of LEAD is dynamic workflow orchestration and data management in a web services framework.

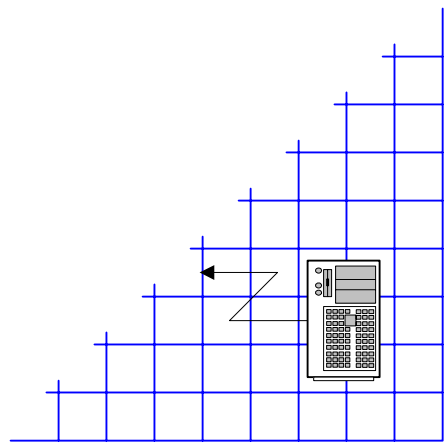
Metadata and myLEAD in the LEAD Project

- Metadata is communicated using the LEAD metadata schema – a profile of the FGDC schema.
- As experiments are run, metadata regarding inputs, outputs, and the processing itself are cataloged and updated.
- Builds on OGSA-DAI by adding activities that use the LEAD schema.



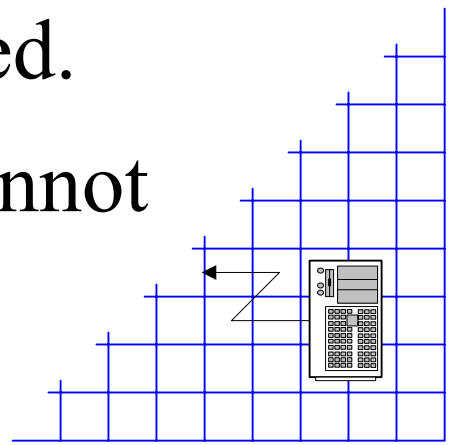
Communicating Metadata in a Grid Environment

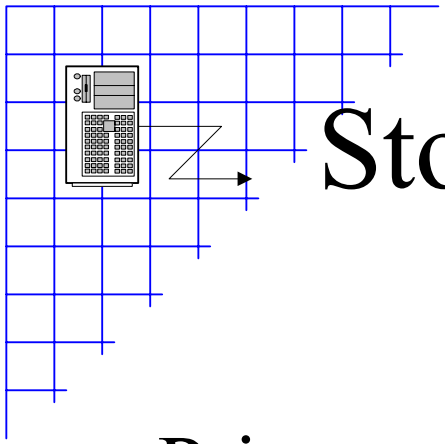
- One or more XML schemas define the metadata for a single data product.
- Schemas are agreed within a community but could vary between communities.
- Hierarchical aggregations of data – often vary by domain.
- Applies to domains with binary data products – biology / bioinformatics differs



Metadata Characteristics

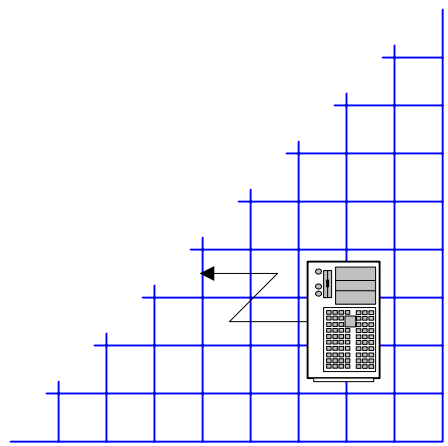
- Queries over metadata attributes of objects (files or aggregates). Queries search for files with the desired attributes.
- Unordered queries – order of attributes in schema are generally not relevant to queries.
- Ordered schema valid results returned.
- Dynamic Attributes – the schema cannot capture all relevant properties.





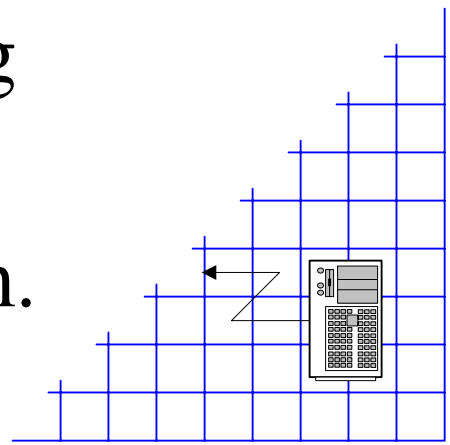
Storing and Querying Metadata
XML vs. Relational

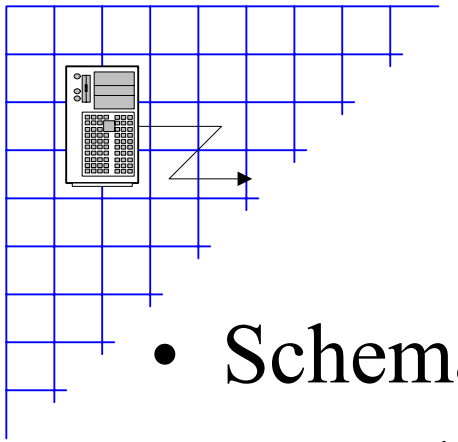
- Prior work we had done benchmarking an XML database versus a relational database showed the XML database to be considerably slower.
- XQuery is a recent standard and not fully supported yet – still XPath.
- XQuery does not address updates.



Approaches to XML Storage
CLOBs versus "Shredding"

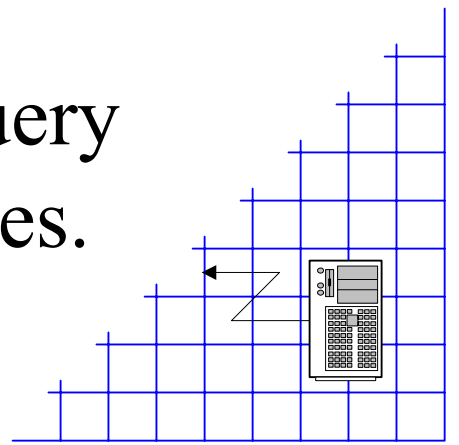
- If the document will never be updated (no changes such as experiment status or additional annotations) then the metadata document can be stored as a Character Large Object (CLOB).
- Shredding creates relational tables to store the XML - usually using an inlining approach.
- Third possibility – Hybrid approach.



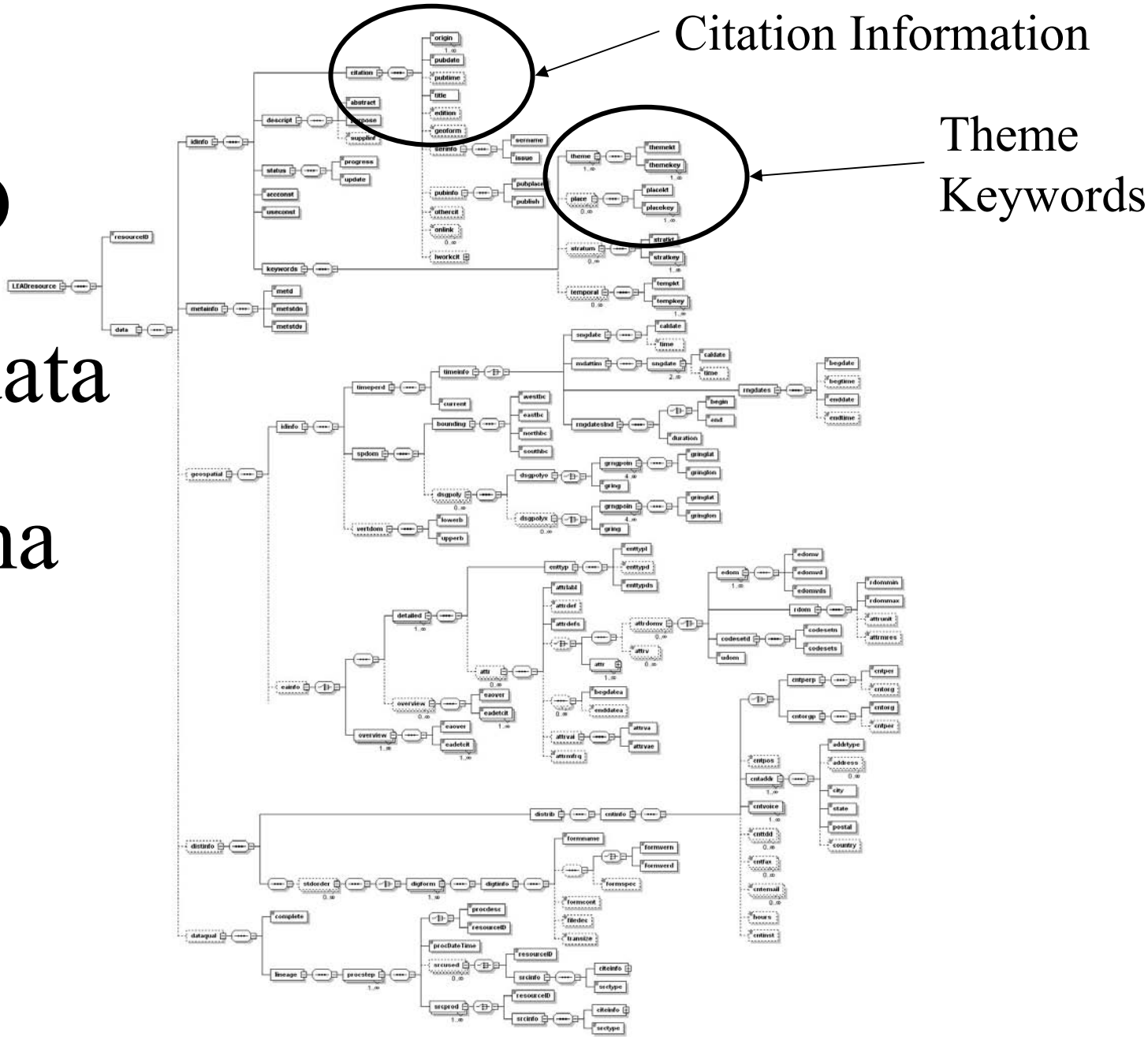


Inlining Approach

- Schema elements inlined where possible
  - Cardinality not greater than one
  - Recursive elements are separate relations
- Ordering added and maintained separately
- Lossless Shredding
  - Nothing omitted, nothing added
- Optimization of inlining through query analysis on weighted targeted queries.



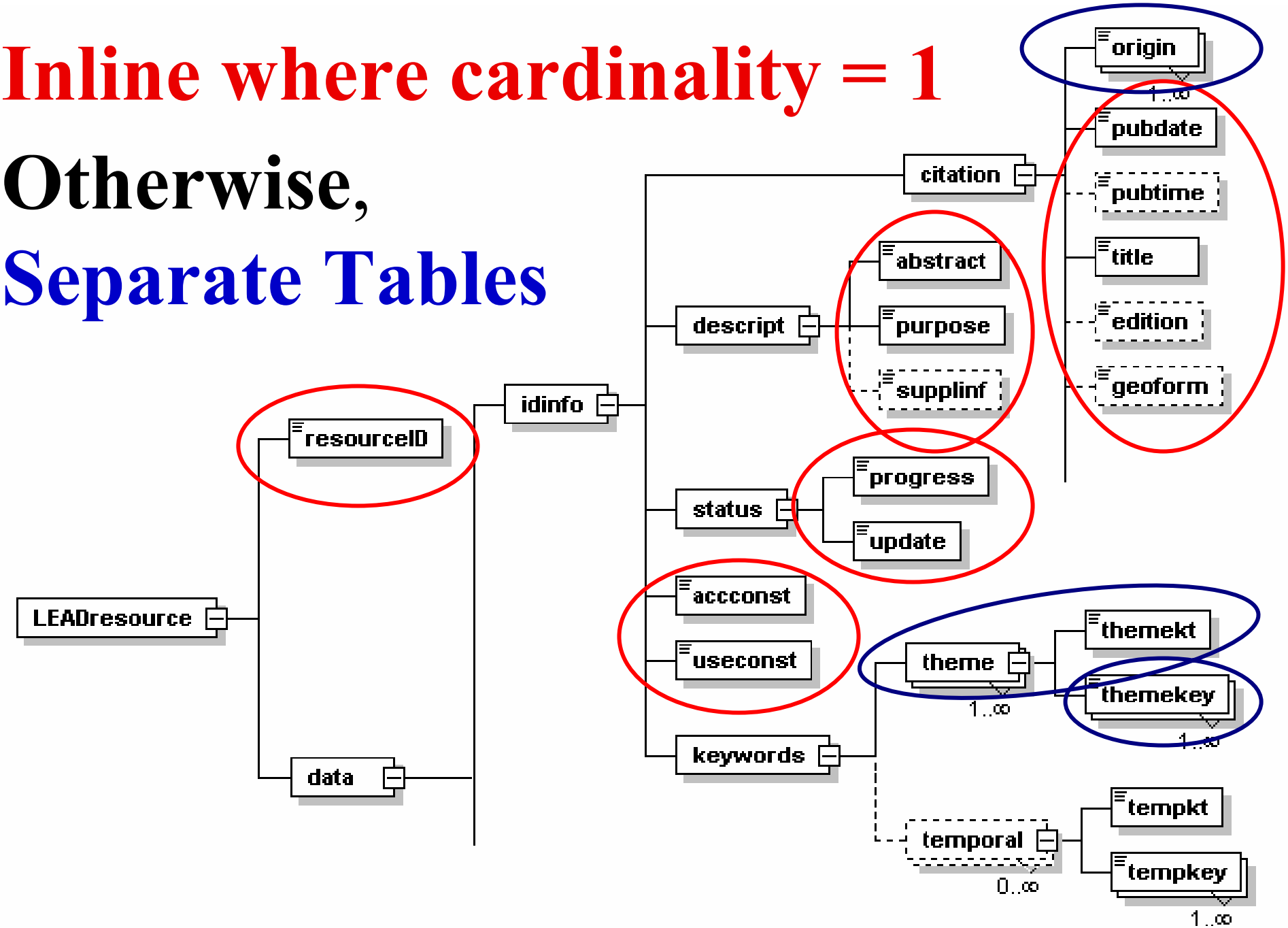
LEAD Metadata Schema



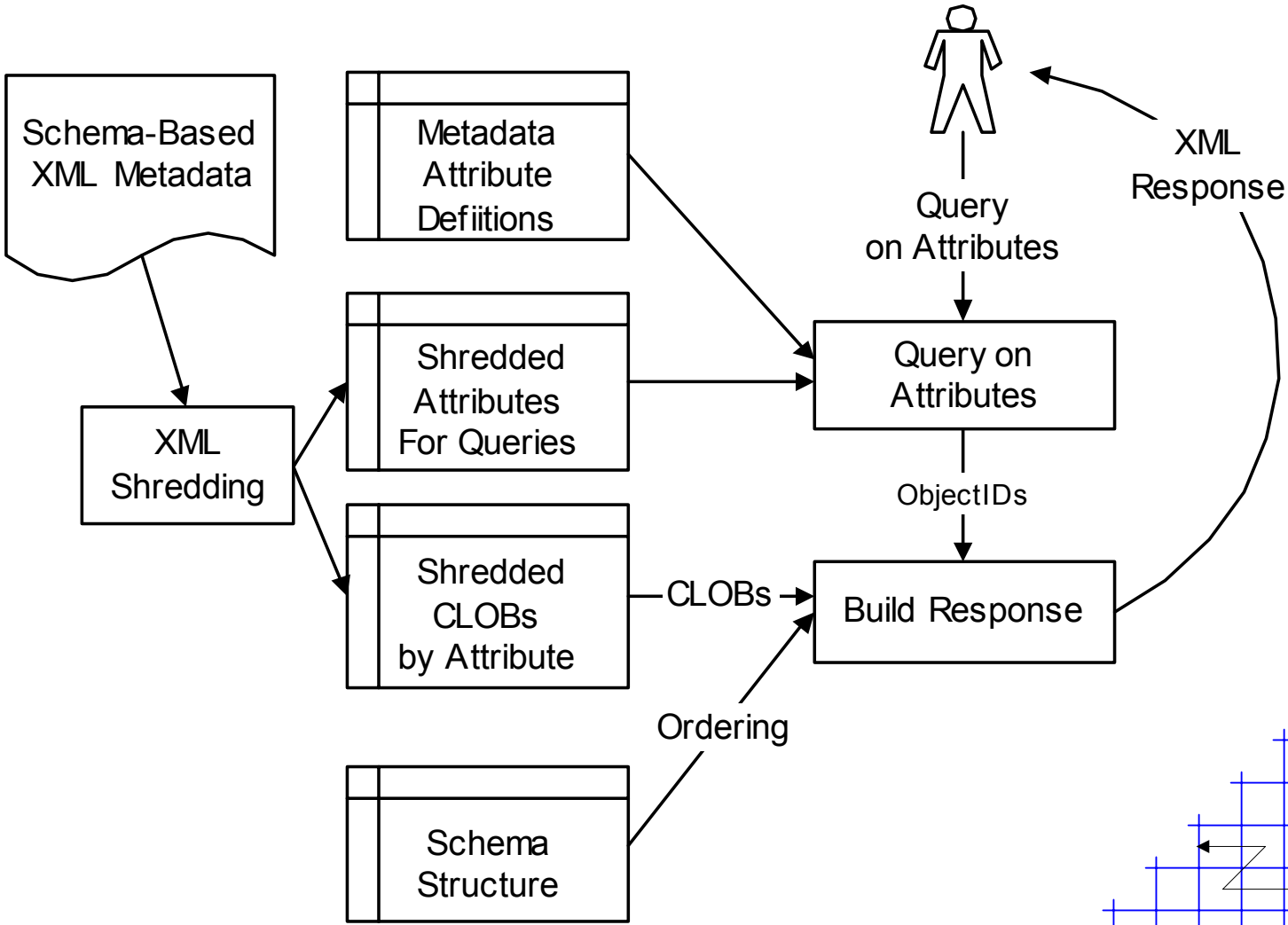
Inline where cardinality = 1

Otherwise,

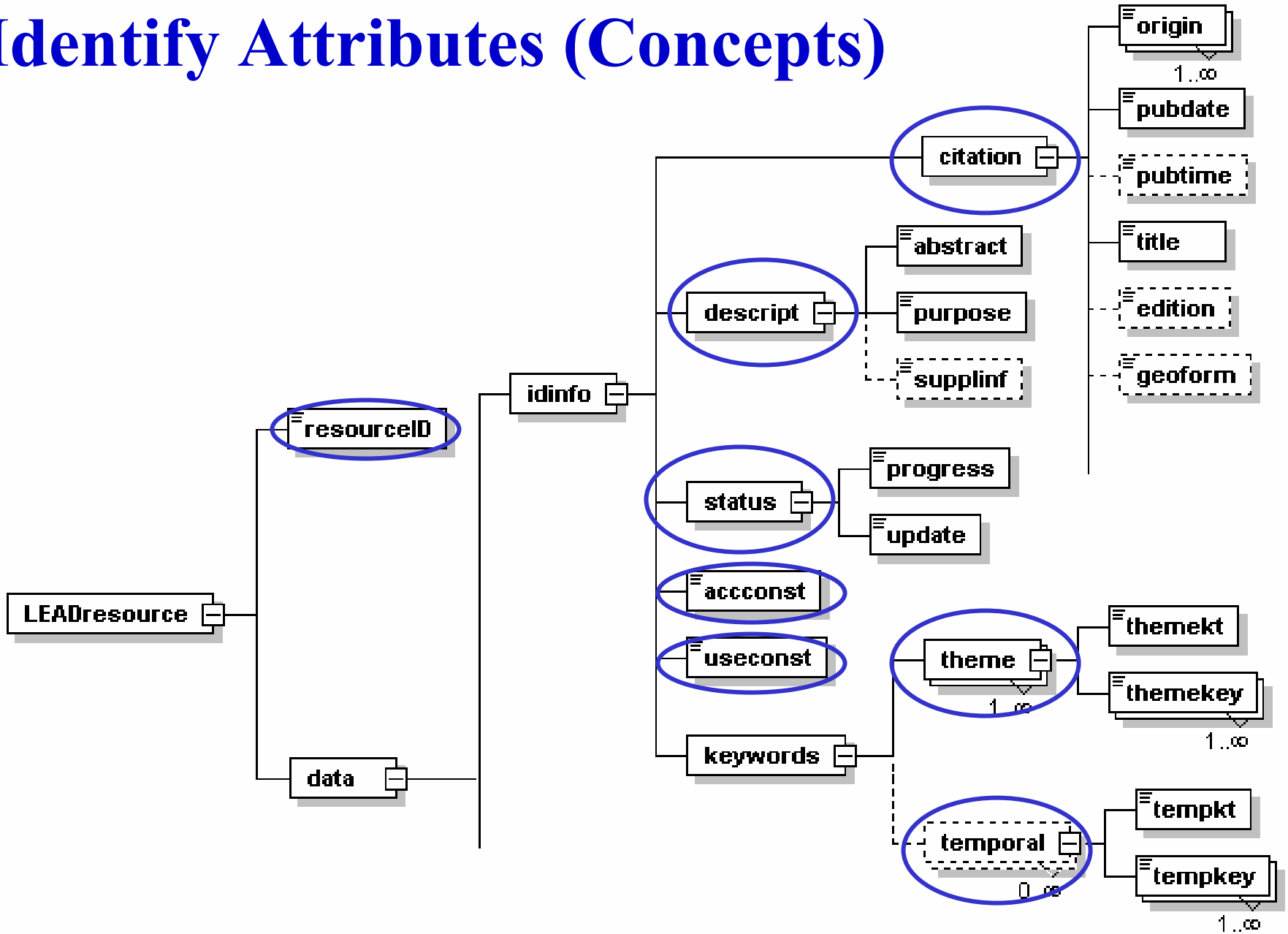
Separate Tables

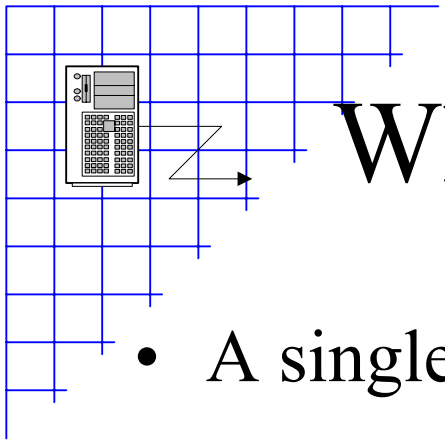


Hybrid CLOB/Shredding



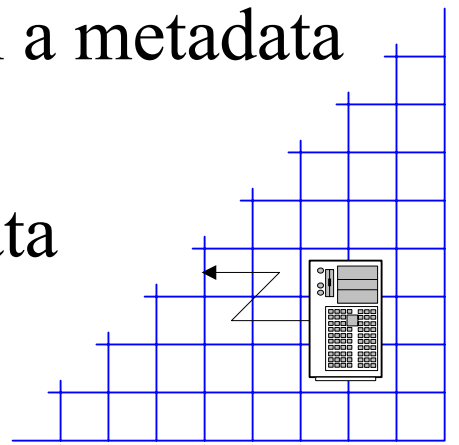
Identify Attributes (Concepts)





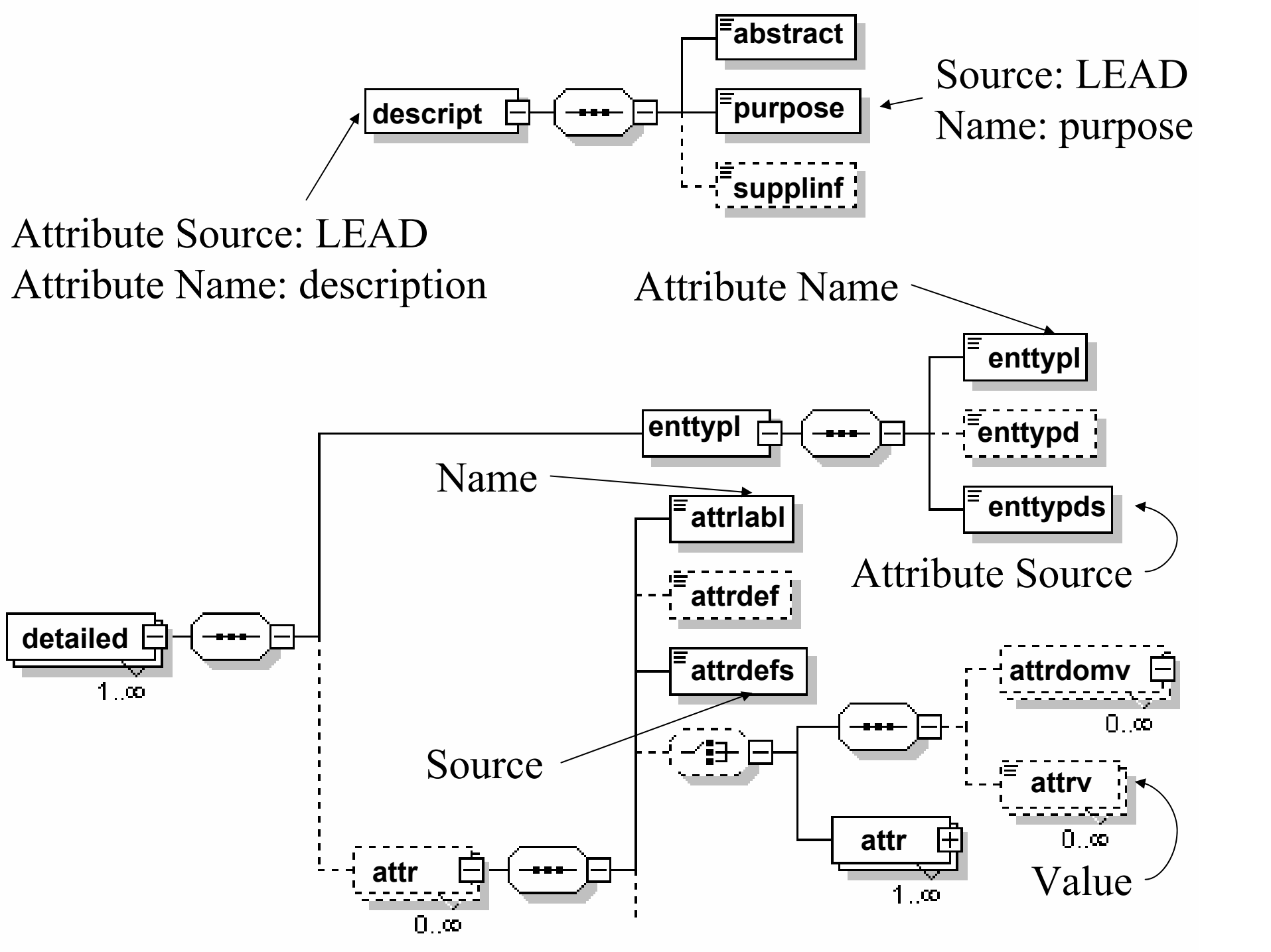
What Is A Metadata Attribute?

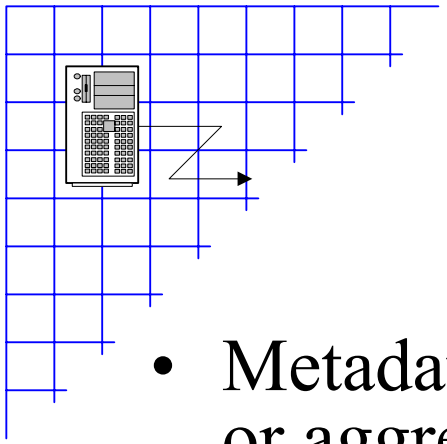
- A single concept (e.g., status)
- Could be a complex concept with a hierarchy.
- Schema elements that allow cardinality > 1 are metadata attributes or sub-attributes.
- Schema elements with XML attributes must be metadata attributes or sub-attributes.
- Any recurrence must be contained within a metadata attribute.
- Every leaf must be contained in a metadata attributed.



Dynamic Attributes

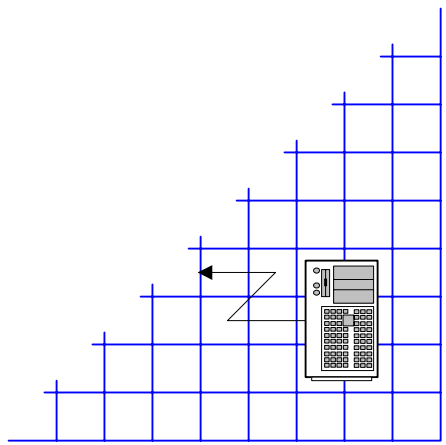
- Structural vs. dynamic metadata attributes
  - Descript node is structural
  - Detailed node is dynamic
- Much of the metadata is model parameters or derived results. Not defined in the schema.
- Data types can be defined in hybrid approach.
- Consistent but expandable set of metadata attributes must be defined for sharing and connecting to an ontology.


abstract
descript
purpose
supplinf
Source: LEAD
Name: purpose
Attribute Source: LEAD
Attribute Name: description
Attribute Name
enttypl
enttypl
enttypd
enttypds
Name
attrlabl
attrdef
Attribute Source
detailed
1..∞
attrdefs
attrdomv
0..∞
Source
attrv
0..∞
attr
attr
1..∞
Value
0..∞



Queries Over Metadata Attributes

- Metadata catalog stores properties regarding files or aggregate (collections, experiments, etc.) objects.
- Scientists are searching for those objects that have certain properties.
- Possibly similar to search engine:
  - Give me the ten files that match the most criteria.
- Results needed:
  - Full metadata for matching objects
  - Object IDs (used to retrieve objects)
  - Selected metadata attributes



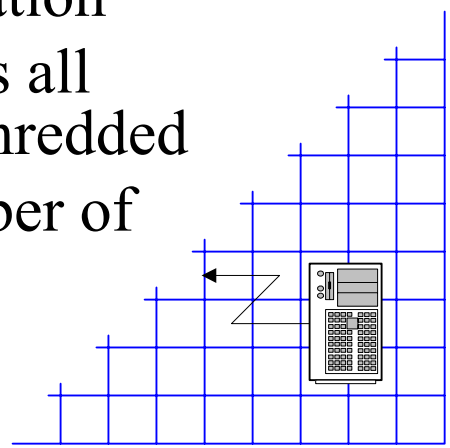
Comparison

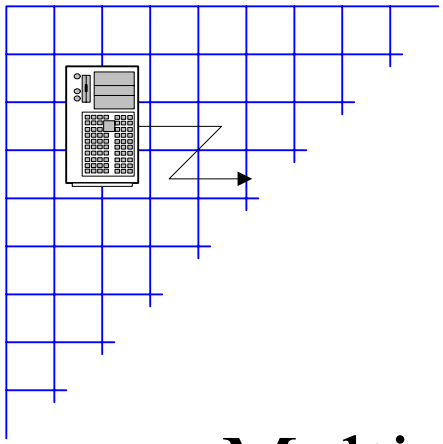
Hybrid

- No temp table needed for CLOBs
- Global ordering
- Could have multiple schemas
- Can be adapted without database changes
- Can be applied to different domains without DB changes
- Not all elements need to be shredded
- Exploits characteristics of metadata catalogs

Inlining

- Temp tables for subqueries used in sorted outer union
- Global order must be updated – Keywords
- Single schema
- Schema change requires DB change
- Must be optimized for each implementation
- Lossless requires all elements to be shredded
- Significant number of joins for results





Going Forward

- Multiple Schemas
- Metadata Catalog Based on Annotated Schema(s)
  - Create XSLT Shredding Based on Schema
  - Populate Metadata Attribute and Element Definition Tables Based on Schema