National Geospatial Digital Archive

Greg Janée

University of California at Santa Barbara

A misadventure in preservation

- 1976
  - Viking probes go to Mars
  - soil data is analyzed for evidence of life
- 1999
  - USC neurobiologist Joseph Miller asks for data
  - NASA has data on tape!
- But…
  - tapes coded “in a format so old that the programmers who knew it had died”

Paradox of preservation

- Is the data valuable?
  - yes: had to travel to another planet to get it

- Is the data being used?
  - no
  - perhaps never again

- How much am I willing to pay for its preservation?
  - as close to zero as possible

Is it worth preserving?

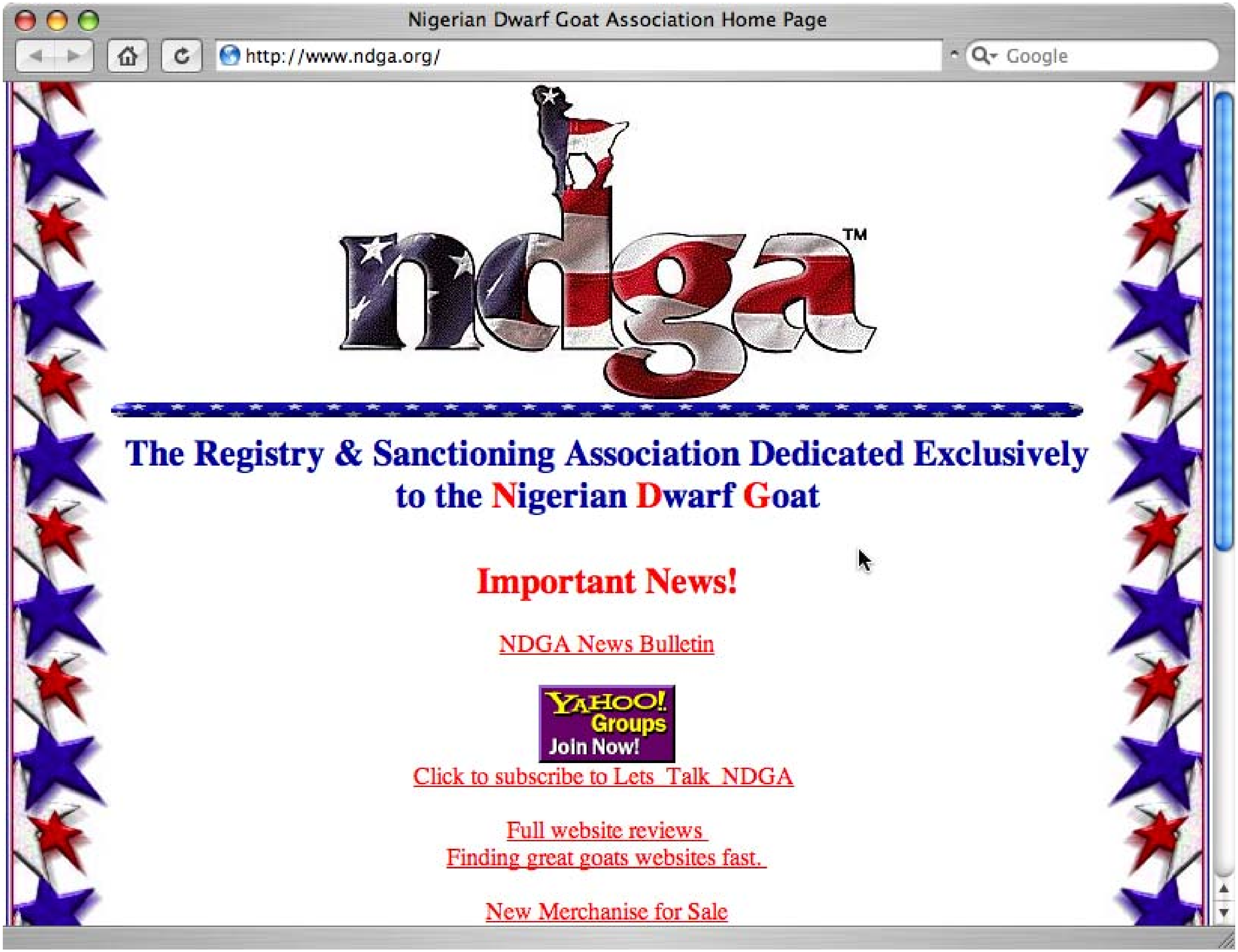
- Keith's equation[*]:
  - (current value) = (intrinsic value) - (cost to use)

- Greg's equation:
  - item is worth preserving for time duration T if:
    - (intrinsic value) * Prob_T(usage) > Σ_T(preservation costs) + (cost to use)

[*] apologies to Keith Johnson, Stanford libraries

Project genesis

- NDIIPP
  - Library of Congress, 2000
  - $100M
  - http://www.digitalpreservation.gov/
- NGDA
  - UCSB (MIL) & Stanford (Branner Library)
  - $2.6M, 3 years
  - geospatial data
  - http://www.ngda.org/



The Registry & Sanctioning Association Dedicated Exclusively to the Nigerian Dwarf Goat

Important News!

NDGA News Bulletin

Click to subscribe to Lets Talk NDGA

Full website reviews
Finding great goats websites fast.

New Merchanise for Sale

Project goal

- *"How can we preserve geospatial data on a national scale and make it available to future generations?"*

- No focus on a particular collection

- Geospatial data
  - discrete chunks
  - relatively highly-structured, well-defined
  - but 90% of our work is generic

Idea #1

- Archival has to be cheap & easy
  - must be distributed
  - little incentive, no funding
  - not sexy

NGDA approach

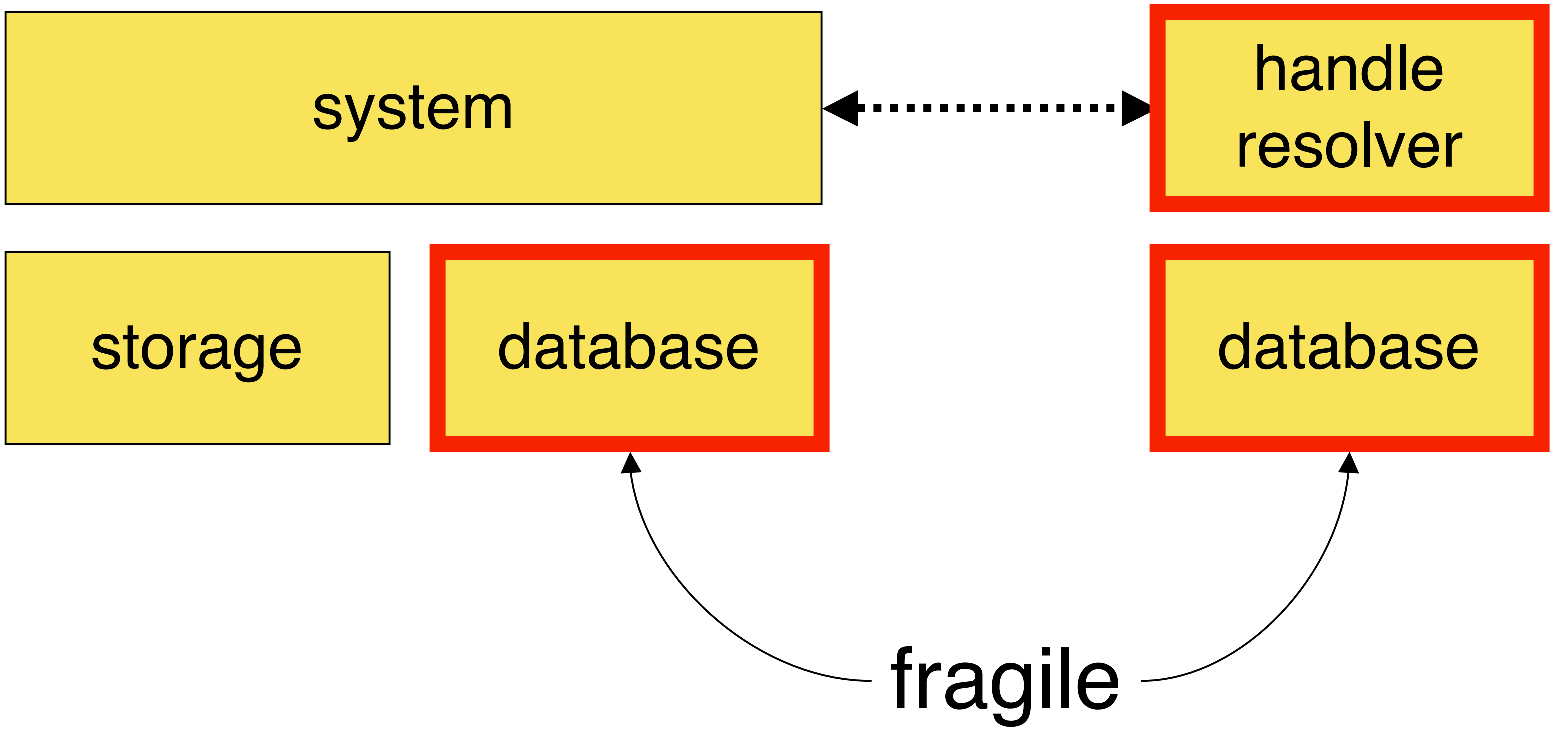
- Compromise: define cheap archive
  - fundamental approach: preservation by co-archival of object semantics
  - ingest: one step up from crawling
  - web access
  - notable for what's missing: discovery, usability

- Foundation for additional functionality
  - e.g., migration
  - prototype archives will offer ADL, OAI access

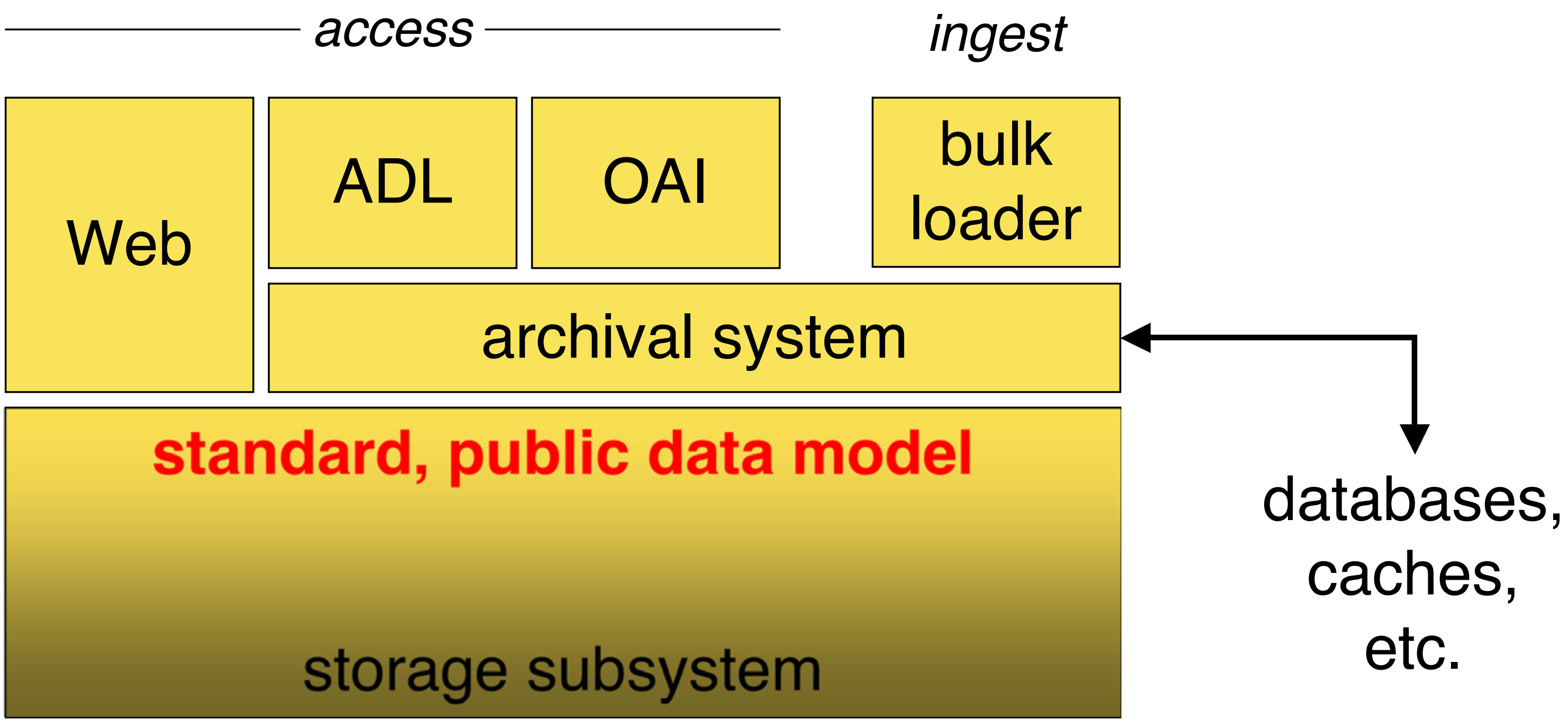
Idea #2

- Archival systems must be designed with their own demise in mind
  - archival *objects* will long outlive any *system* that manages them
  - system-level migrations will occur
  - at inopportune times

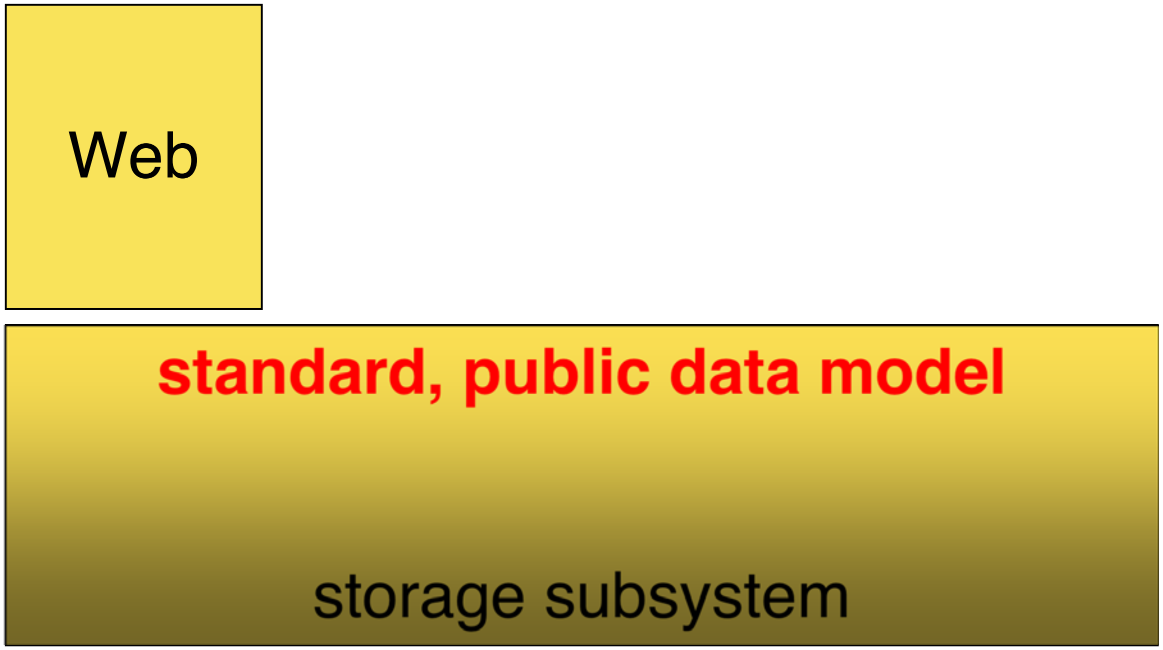
Typical repository architecture



NGDA architecture



Post-NGDA architecture

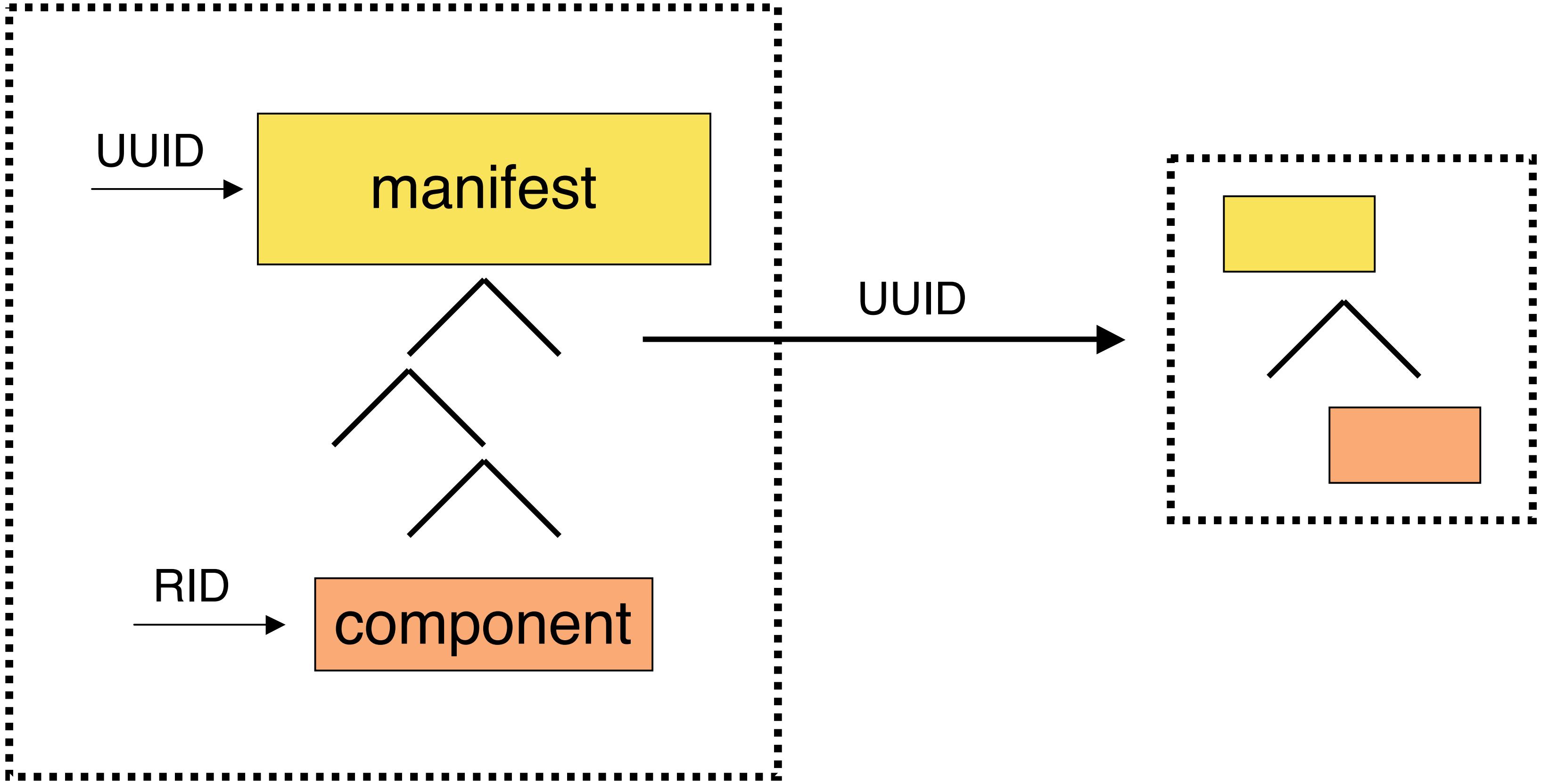


Storage system requirements

- Req's:
  - associate UUIDs/RIDs with bitstreams
  - retrieve global/local bitstream by UUID/RID
  - determine (parent) UUID of any bitstream
  - list all UUIDs

- Satisfied by:
  - any filesystem
  - any kind of UUIDs
    - tag:library.ucsb.edu,2005:*identifier*

Archival objects



Archival object representation

- Components are files
- Manifest is an XML document

- Other approaches
  - OAIS: archival information packages (AIPs)
  - XMLtape

Ingest

- Ingest template defines
  - common structure of objects to be ingested
  - necessary validations
  - associations to other objects
    - assumes pre-loading of semantic definitions
  - policies, rights, etc.

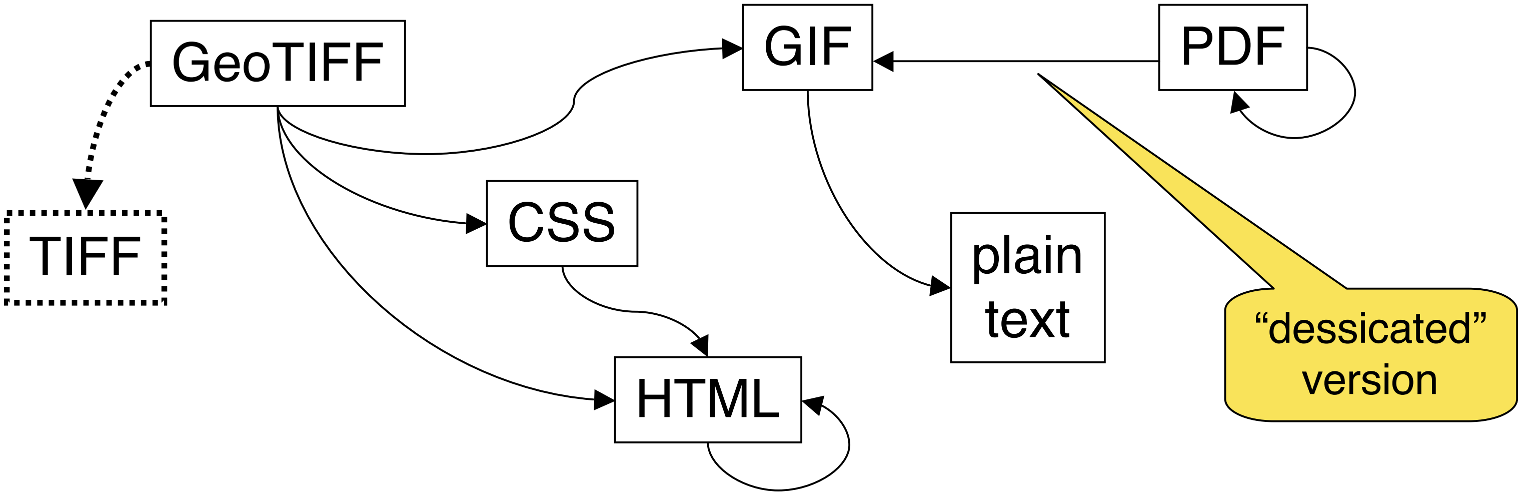
- Represents choke point
  - requires human evaluation

Format registry

- We’re developing one
  - who isn’t?
- Serves as archive of format specifications
- How broadly to interpret “format”?
  - traditional file format
  - product
  - series, collection, arbitrary set

Format dependencies

- Consider dependency graph induced by format specifications
- Def: a format is recoverable if the format of its specification is recoverable
- Axioms: plain text, HTML are recoverable



Challenges

- Making ingest easy, easier, easier-er, ...
- GIS formats
  - very complex: topology, layer, coverage, project
  - proprietary
- MODIS
  - multiple petabytes
  - format (HDF) is not well-defined
  - moving to on-demand computation of products
  - lineage important
  - copious additional semantics

Misadventure, redux

- What if there had been an NGDA-like solution?
  - format specification would have been archived
- Limitations
  - data not necessarily immediately usable
  - format specification itself not necessarily viewable
- But limitations can be addressed according to usage, available resources

Questions for you

- Archival systems
  - definition? functionality?
- Storage systems
  - definition? functionality?
- Archival object representation
  - discrete files vs. AIPs?
- GIS formats
  - “dessicated” form?