Slide 1

Welcome to the Mad Science of Search

Definitions:

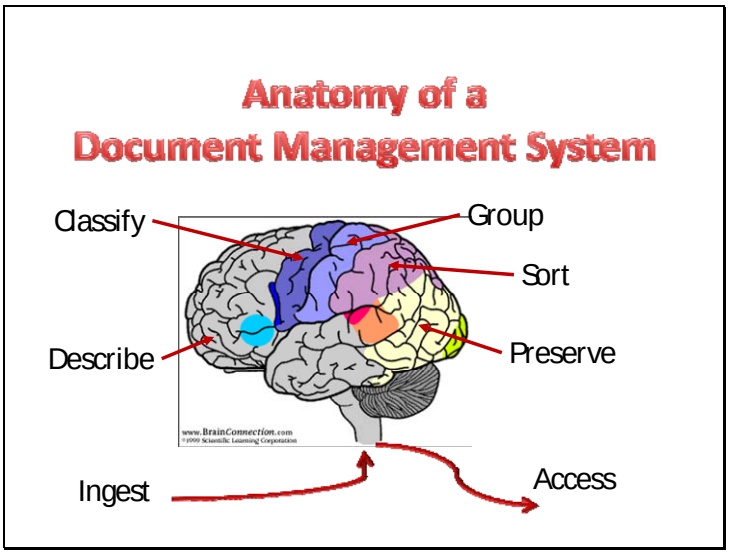
Document, Data, Record, Object

What to call Digital Assets is a hotly debated topic. In this presentation these terms are used interchangeably to reference Digital Assets.

Metadata, Metadata Field, Markup

Throughout this presentation metadata references imply the use of a database, however it could also be markup within an object, properties attributed to a system folder, html tags or any other type of method that links descriptive data with an object. The implied use of a database simply comes from my background and I am completely open to the use of other methods. ☺

Slide 2

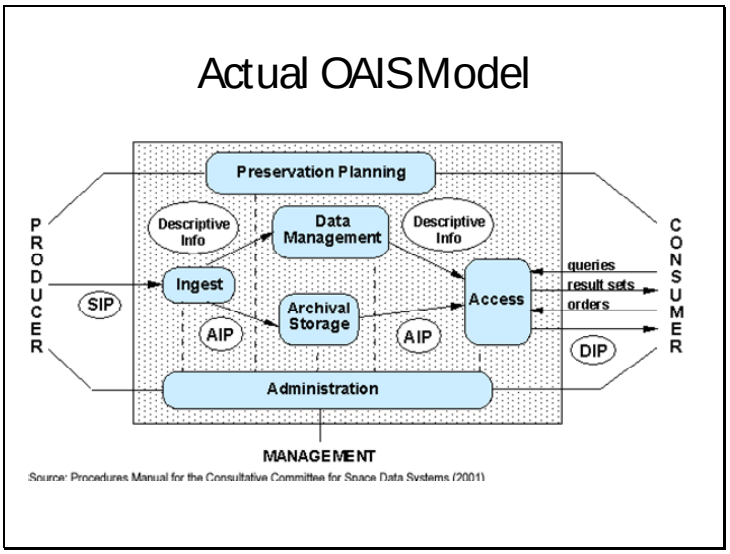


Anatomy of a Content/Document Management System

- Ingest Data
- Describe Data
- Classify Data
- Group Data
- Sort Data
- Preserve Data
- Access Data

This is the OAIS Model (Open Archival Information System), ISO Standard 14721:2003

Slide 3

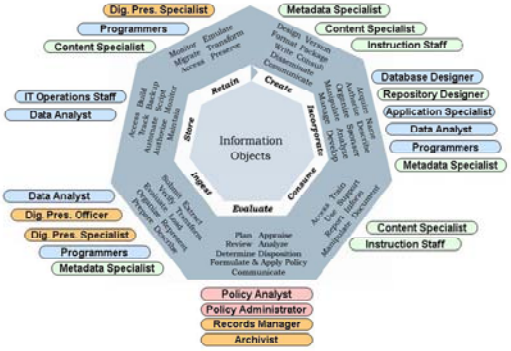


Actual OAIS Model
SIP, AIP, DIP = Information Packets
Submission Information Packet
Archival Information Packet
Descriptive Information Packet

Could be metadata, links to associated data, classification, retention period,etc.
Can change as document moves through its lifecycle

Slide 4

Actual OAIS Model

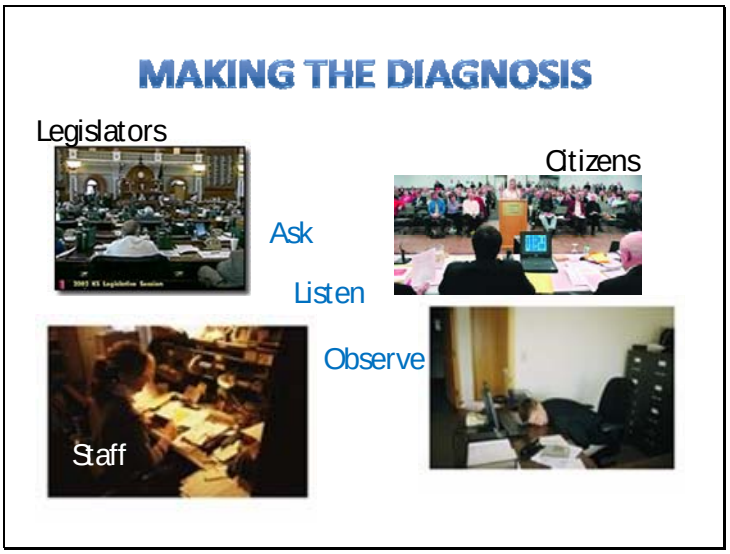


The Functional Dimension of the Information Lifecycle

Functional Dimension of the Information Lifecycle
Note that Retrieval, Extraction, Access to objects is necessary to each user group

Links to ISO Standards and Reference Models:
http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

Slide 5



Making the Diagnosis

- Identify your patients – Legislators, Secretarial, Attorneys, Analysts, Citizens
- Ask them questions – What records, documents, data do they need? How quickly? Is a long, broad results list ok or do they want THE ONE back? Will they learn advanced search techniques?
- Listen to your patients
- Observe their symptoms – What do they really do? How much time do they really have? When they use Google, do they really go through the list or use the first hit, then try different search criteria? What search criteria are they using? (This becomes your metadata)

Slide 6

Research all Factors Before Diagnosing

- Use all available resources – Librarians, Record Officers, etc.

Gather information to answer these questions:

Develop Search

- Determines type of search needed – Boolean characters, full text, metadata, stemming,
- How advanced will search be?
- Can user "drill down" within results list?
- Does search need to return multiple file formats or specific file format for different user groups?
- What security needs to be applied for different user groups?
- Do we need to return document sets or just individual docs? (i.e. group bill, fiscal, explainer, minutes by bill #)
- Does user need to know life expectancy of records they are accessing?
- When record expires, will search update dynamically?
- What are relationships between records? What metadata, file structure, linking, etc. is required to allow user to retrieve associated records?
- What collections will be included in a specific search? Collection size?
- Are records classified by importance, sensitivity, regulations? How will this affect searching return time?

- What is the location of records – online, near-line, tape storage?

User Interface

- Don't display hits in results list user can't access
- User security level is clear – view or edit
- How will user access records over time? (As record moves from active to archived)
- Does user need to know file format or is a link to free readers available? (Adobe, Visio, DIVX, etc.)
- Preview of document
- Design of results list:
    - Display search criteria highlighted
    - Return criteria in context from hit list
    - Ranking – by metadata, title, anchor text, etc.
    - Allow user to customize screen - sort results, number of hits displayed, metadata displayed, etc.
    - Easy instructions for advanced searching

Slide 7

The Cure

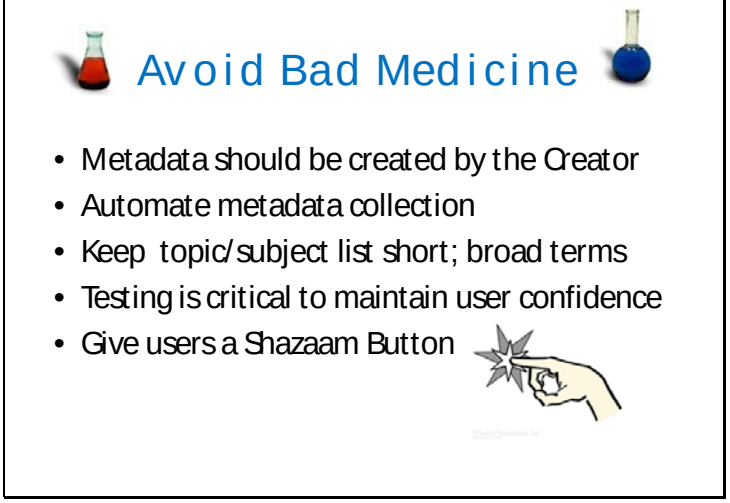
- Metadata
- Full Text
- "Google" Style
- Grouping
- Vertical Search

The Cure

- Metadata – Specific Field Search (Remember the definitions at the beginning!)
  - Search on one or multiple metadata fields
    - Only as effective as the quality of the metadata
    - Can search by topic or subject
  - Full Text – Build index on individual words or phrases within document, use of noise list
    - Can return too many low quality matches
  - "Google" Style – Bayesian Theory of Probability – Returns most likely hits based on past searches, good enough but not exact. It does self-correct as more search data is accumulated. (Also used by spam filters)

    Google combines this method with others including text matching and Page Rank – ranking pages by the number and quality of links to it.
  - Grouping – Set of associated documents are returned. Association can be through storage location, metadata, system level link (attachment)
  - Vertical Search, Defined Collections, Classification – Ask.com

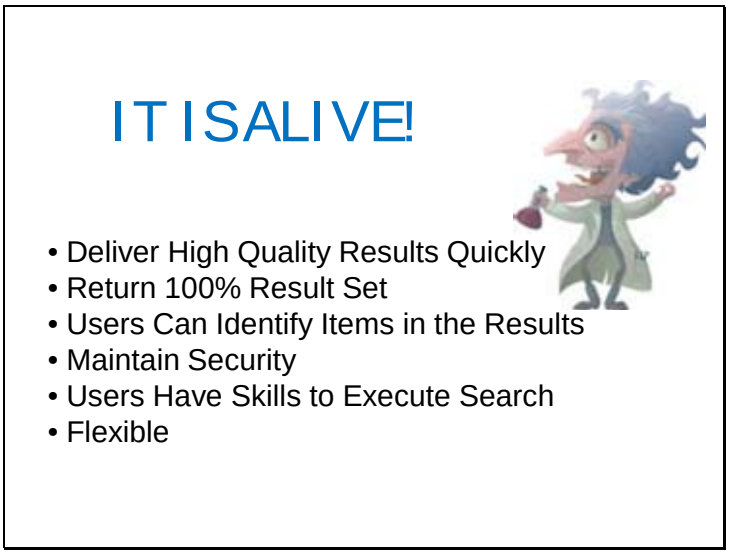
Slide 8



Avoid Bad Medicine

- Metadata should come from creator, not the secretary
- Automate metadata collection as much as possible, otherwise it's too cumbersome and staff won't add it
    - Pull title from document
    - Pull author, date from system data
    - Use rules and masks to define metadata wherever possible
- Keep subject list short – too much to scroll through, users want to add their specialized terms
    - Talk to your legislative librarian about their list
- Test!! Maintain user confidence in results
- Shazaam Button – Find the most used search and make it easy for them!! In Kansas it's the Bills and Associated Docs search - Returns all versions of the bill, committee minutes, fiscal notes, Supp notes, bill explainers, supporting documents reference the correct version of the bill in the metadata

Slide 9



Conclusion – It Is Alive!

- Deliver high quality results quickly – what the user is actually looking for
- Return 100% Result Set – Users must be confident they are retrieving every document that meets the search criteria
- They must be able to quickly identify items in the results list
- Security must be maintained
- Users have necessary skills to execute the search
- Flexible – Must be able to adapt as user needs change

Kansas Information Technology Architecture
http://www.da.ks.gov/itec/documents/kita_ver11_final.pdf