The Case for Adaptive Hierarchical Metadata

Steve Worth

Director, Global Innovation Operations, EMC

Henrique Nogueira

SW Engineer
Eldorado Research Institute

Arthur Guerra

SW Engineer
Eldorado Research Institute

iRODS User Group Meeting 2016
6/9/2016

EMC²

Find out what your DNA says about you and your family.

- Learn what percent of your DNA is from populations around the world
- Contact your DNA relatives across continents or across the street
- Build your family tree and enhance your experience with relatives

order now $99

5% =

71M =

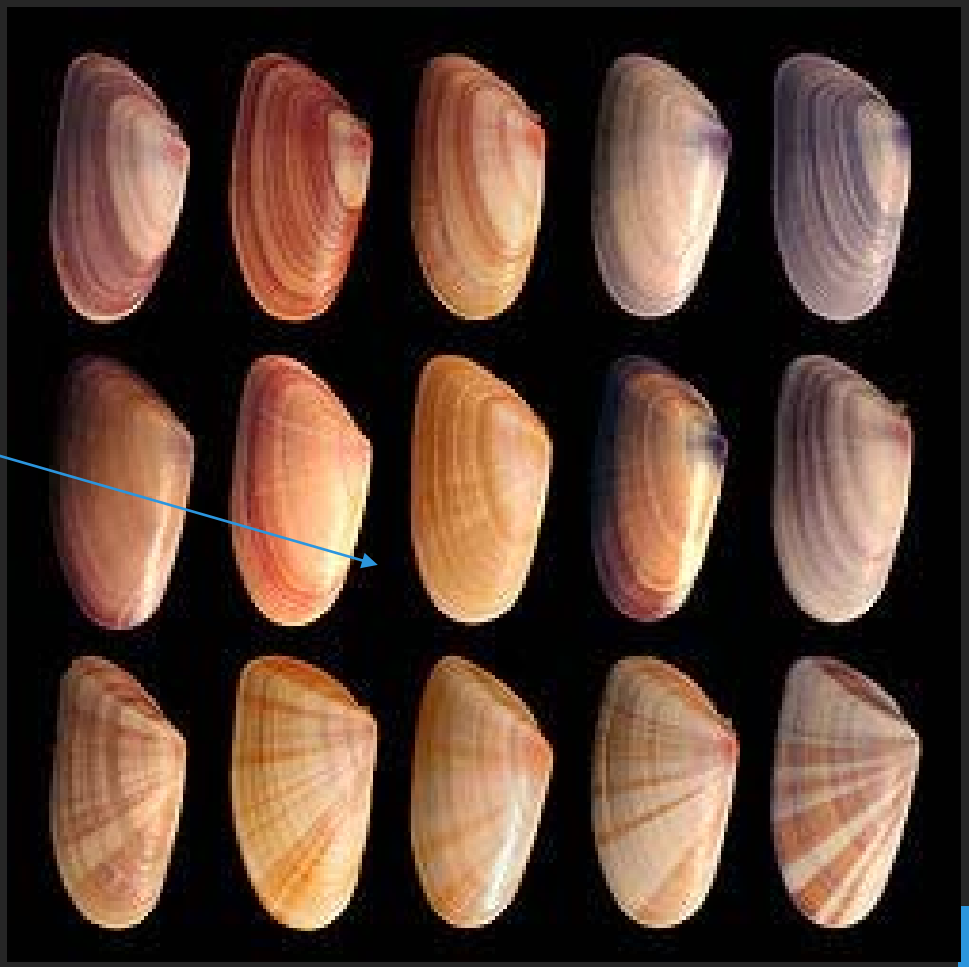
20 ExB =

$2.8 Billion/Yr

```
usxxpaeglem1:ino paegle$ ls -l
total 0
-rw-r--r--  1 paegle  CORP\Domain Users  0 Sep 23 12:29 my.bam
-rw-r--r--  1 paegle  CORP\Domain Users  0 Sep 23 12:29 my.log
usxxpaeglem1:ino paegle$
```


EMC²

STUDY_ID: NIH2015-UO-0001456
POPULATION: 185
SUBJECT_BASE_TYPE: Shell
SUBJECT_TYPE: Mollusk
BASE_COLOR: 15
PATTERN_TYPE: Radial



Coping Behaviors

Eliminate data

Pay per use (copies / transfer / or both)

Keep minimal copies (one)

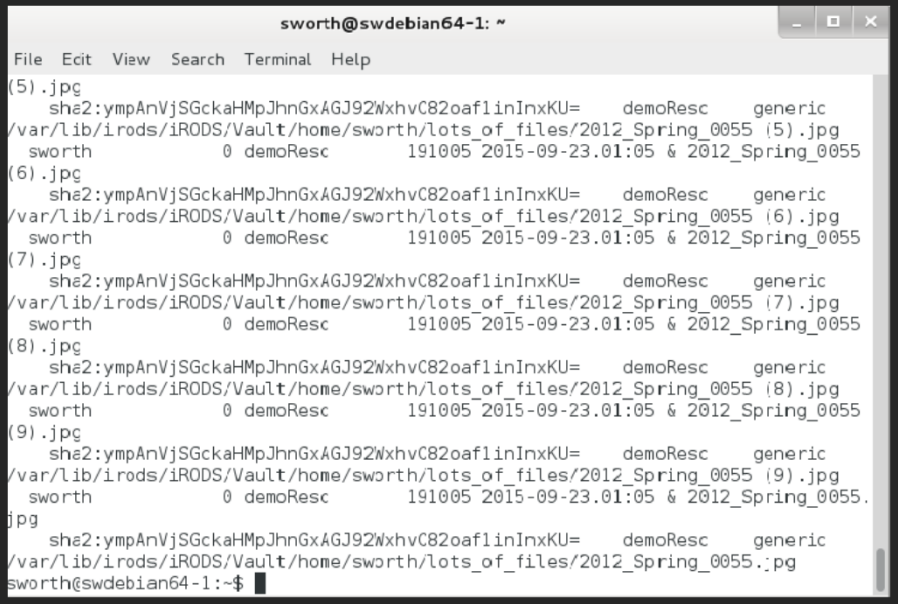
Advanced Metadata is Critical

- Metadata can be stored/transported more cheaply
  - An object with 1000 tags, 4096 character limit, 3 tag AVU model requires < 12MB for the tag set.
- Maximize storage assets - find what's valuable, no matter where it's located
  - Eliminate data that is not valuable (so called "dark data")
- Automate movement and processing of data
- Securely share data with collaborators
  - Easier to separate PHI from non-PHI data
  - Can be used to ensure data is authentic and unaltered
- Metadata supports indexing / fast search methods
  - Ideal for non-SQL techniques

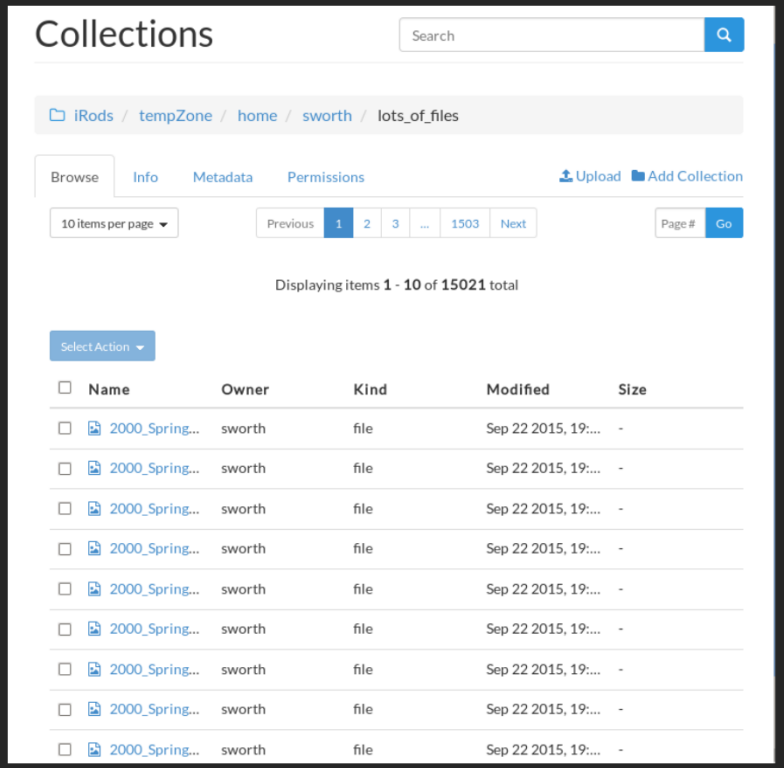
Working With iRODS…

- **User commands** (subset)
  - iinit
  - iput
  - iget
  - imkdir
  - ichmod
  - icp
  - irm
  - ils
  - ipwd
  - icd
  - irepl
  - iexit
  - ipasswd
  - ichksum
  - imv
  - iphymv
  - ireg
  - irmtrash
  - irsync
- **Iadmin commands** (subset)
  - lt
  - lr
  - ls
  - lz
  - lg
  - mkuser
  - moduser
  - aua
  - rua
  - rpp
  - rmuser
  - mkdir
  - rmdir
  - mkresc
  - modresc
  - modrescdatapaths
  - rmresc
  - mkzone
  - modzone

MetaLnx – iRODS Simplified

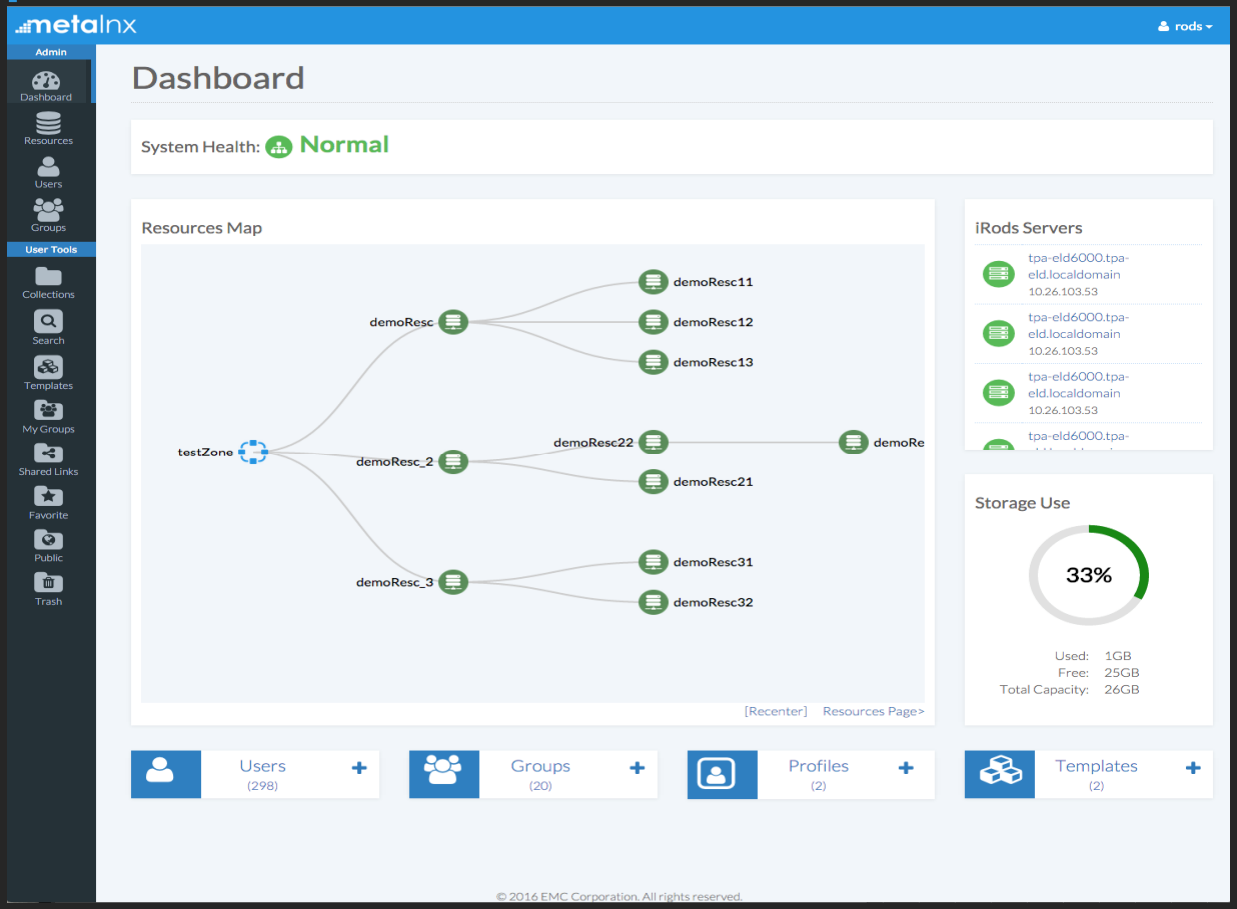


Complex → Organized

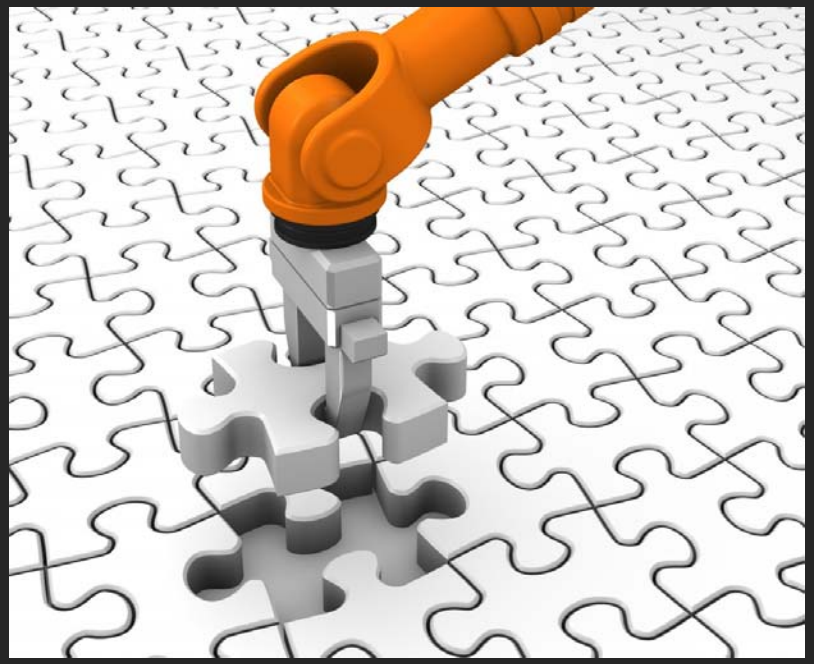


EMC²

Simplified Admin



Metadata Automation



1. FASTQ_FILE, path to fastq file on ftp site
2. MD5, md5sum of file
3. RUN_ID, SRA/ERA run accession
4. STUDY_ID, SRA/ERA study accession
5. STUDY_NAME, Name of stury
6. CENTER_NAME, Submission centre name
7. SUBMISSION_ID, SRA/ERA submission accession
8. SUBMISSION_DATE, Date sequence submitted, YYYY-MM-DAY
9. SAMPLE_ID, SRA/ERA sample accession
10. SAMPLE_NAME, Sample name
11. POPULATION, Sample population, this is a 3 letter code and it is defined in README.populations
12. EXPERIMENT_ID, Experiment accession
13. INSTRUMENT_PLATFORM, Type of sequencing machine
14. INSTRUMENT_MODEL, Model of sequencing machine
15. LIBRARY_NAME, Library name
16. RUN_NAME, Name of machine run
17. RUN_BLOCK_NAME, Name of machine run sector (This is no longer recorded so this column is entirely null, it was left in so as not to disrupt existing sequence index parsers)
18. INSERT_SIZE, Submitter specifed insert size
19. LIBRARY_LAYOUT, Library layout, this can be either PAIRED or SINGLE
20. PAIRED_FASTQ, Name of mate pair file if exists (Runs with failed mates will have a library layout of PAIRED but no paired fastq file)
21. WITHDRAWN, 0/1 to indicate if the file has been withdrawn, only present if a file has been withdrawn
22. WITHDRAWN_DATE This is generally the date the file is generated on
23. COMMENT, comment about reason for withdrawal
24. READ_COUNT, read count for the file
25. BASE_COUNT, basepair count for the file
26. ANALYSIS_GROUP, the analysis group of the sequence, this reflects sequencing strategy. Currently this includes low coverage, high coverage, exon targetted and exome to reflect the 2 non low coverage pilot sequencing stratergies and the 2 main project sequencing stratergies used by the 1000 genomes project.

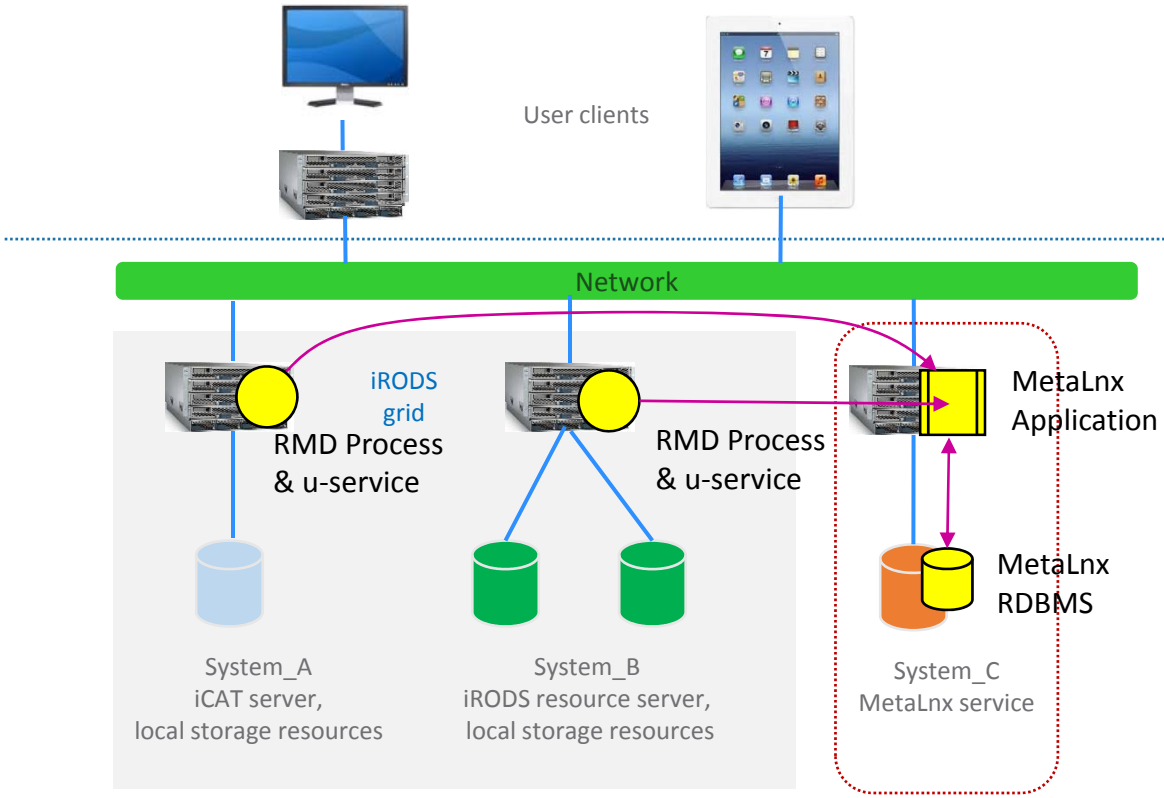
Metadata Templates

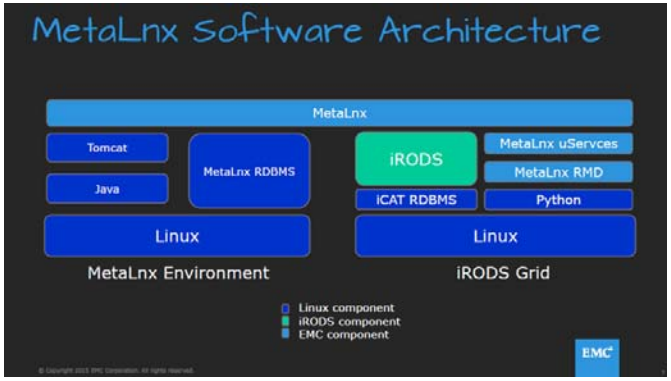
Metadata Searching

MetaLnx Value Proposition:
Simplified Metadata Management

- Simple, easy iRODS grid administration
- Simple, intuitive tool for non-IT researchers:
  - Collection management
  - Automated embedded metadata extraction
  - Metadata searching / editing
  - Metadata templates – consistent tagging of data
- Scale out design supports PCs and tablets
- Extensible to any data vertical

EMC MetaLnx Architecture





EMC iRODS Efforts

- MetaInx
  - An iRODS Administration and Metadata Interface tool
    - Rev 1.0 available on Github at:
      - https://github.com/sgworth/metalnx-web (MetaInx application)
      - https://github.com/sgworth/metalnx-rmd (Remote Monitor Daemon)
      - https://github.com/sgworth/metalnx-msi (iRODS microservices for Genomes)
      - **will be available via www.emccode.com soon**
    - Supports iRODS 4.x
      - Git has instructions & tools to build RPM, DEB, and Docker images
    - Please try. If you like join the community!
- Storage Resource Drivers for iRODS
  - Isilon – HDFS interface
  - ECS – Atmos object interface
    - Will be posted this summer

EMC²

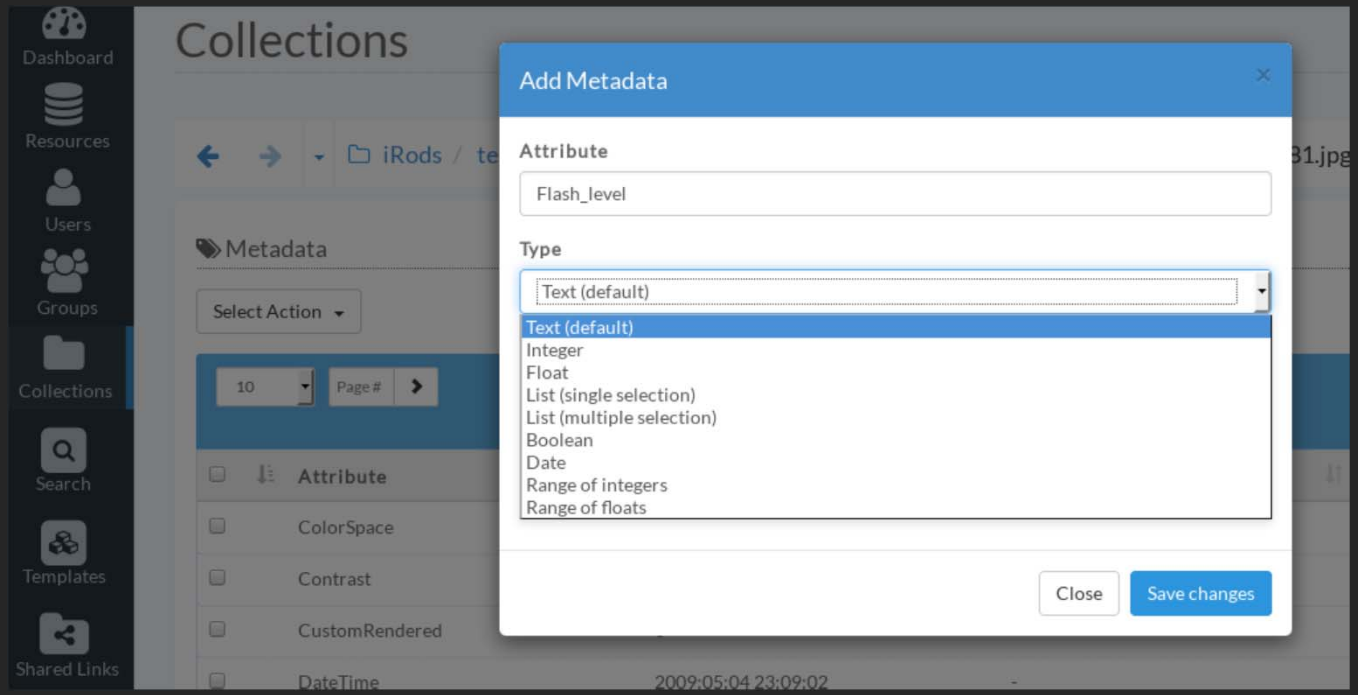
Roadmap

- Metadata Validation
- Turnkey rule set deployment
- Self registration
- Faster, large file transfer support
- Federation/Quota support
- Better data in place linking

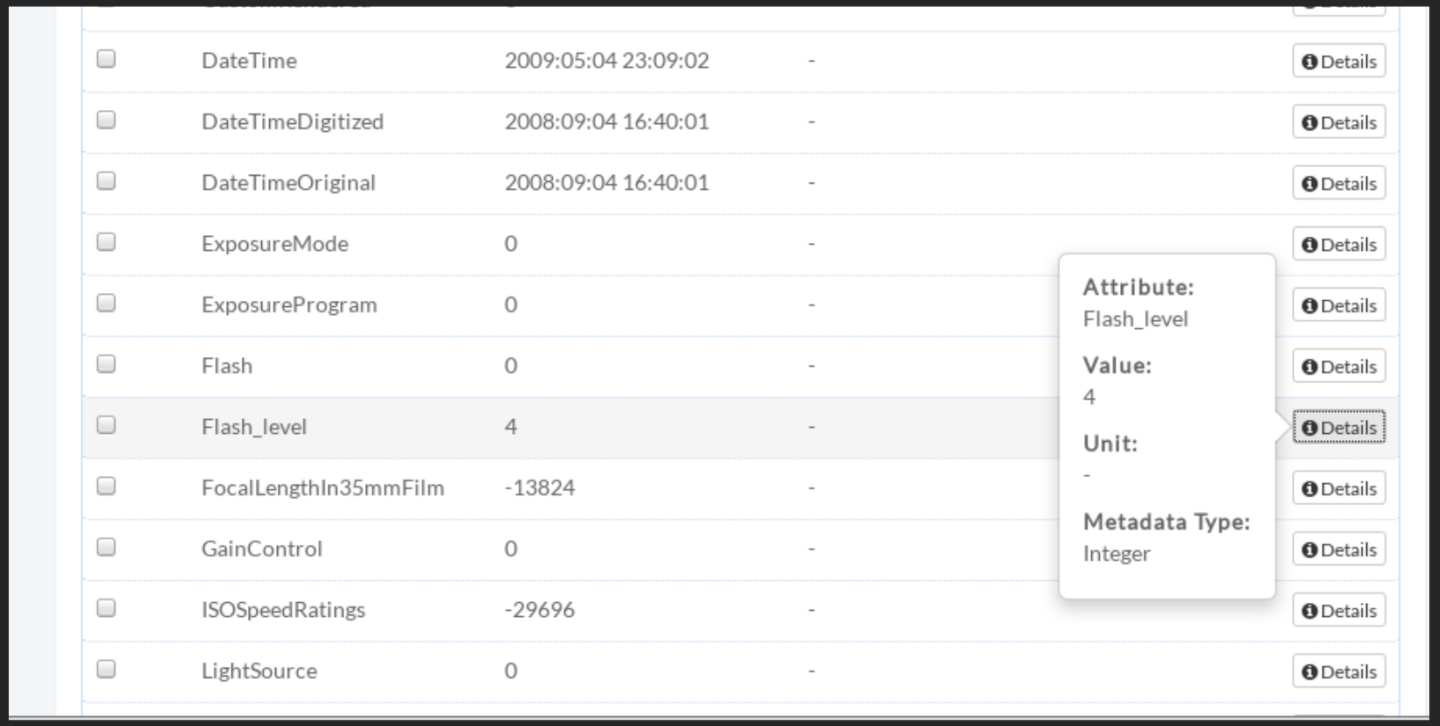
Data Validation – Two Models

- "Lazy" Validation
  - Validation can be added (or not) for each AVU added on each object.
  - Enforced through the UI
  - Can be used with Metadata templates
  - Type options:
    - String (default)
    - Integer
    - Float
    - List (single selection)
    - List (multiple selection)
    - Boolean
    - Date
    - List of Integers
    - List of Float
- Strong Validation
  - Attribute names and validation rules defined in a system vocabulary
  - Users can also have personal vocabularies
  - System settings, forced via UI for "none, system first, user first, user selection" Also can lock system.
  - In "locked" mode an AVU must come from the vocabulary (per rule) and the value must match the validation rule
  - Options – same as "Lazy" validation

Validation Example



Validation Example



☐	DateTime	2009:05:04 23:09:02	-	Details
☐	DateTimeDigitized	2008:09:04 16:40:01	-	Details
☐	DateTimeOriginal	2008:09:04 16:40:01	-	Details
☐	ExposureMode	0	-	Details
☐	ExposureProgram	0	-	Details
☐	Flash	0	-	Details
☐	Flash_level	4	-	Details
☐	FocalLengthIn35mmFilm	-13824	-	Details
☐	GainControl	0	-	Details
☐	ISOSpeedRatings	-29696	-	Details
☐	LightSource	0	-	Details

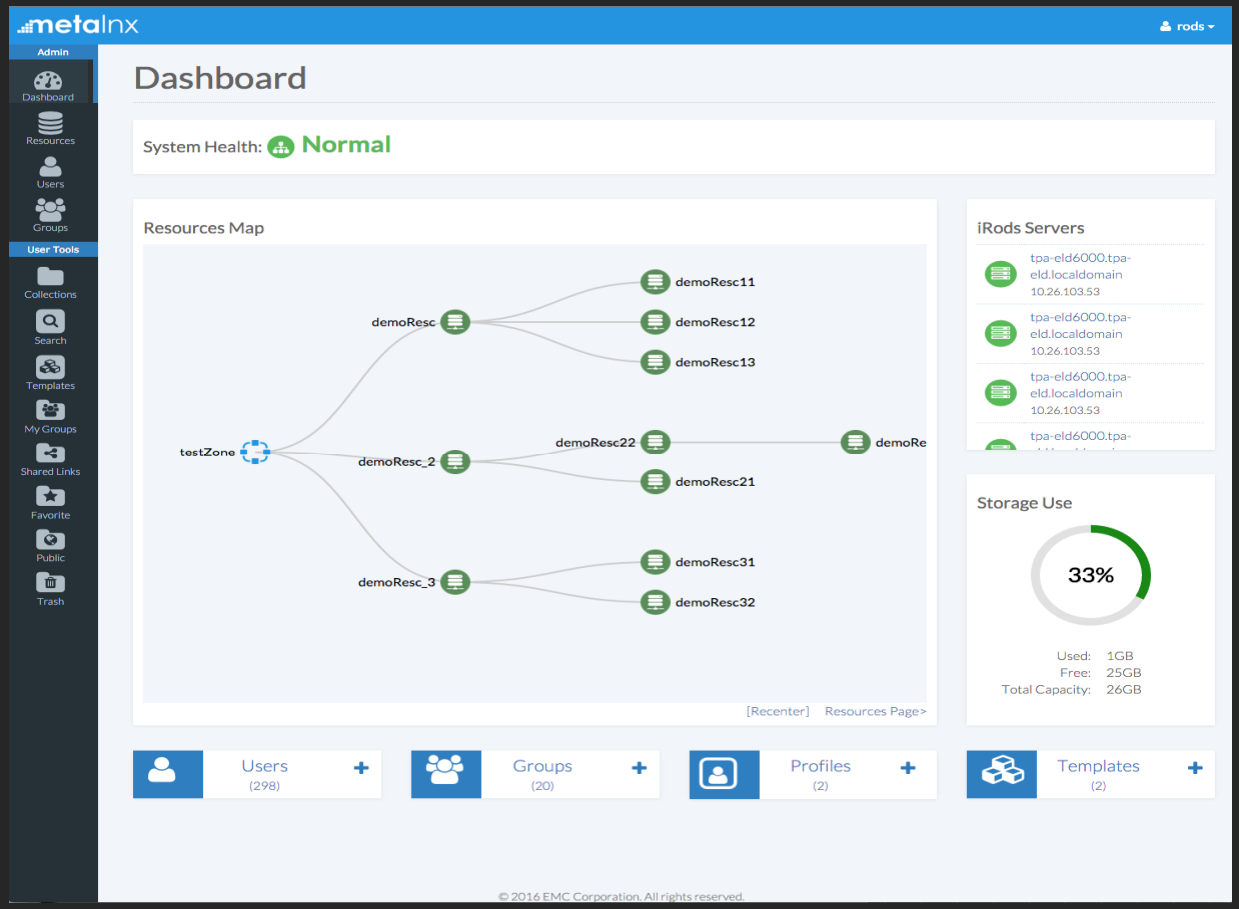
Attribute:
Flash_level

Value:
4

Unit:
-

Metadata Type:
Integer

MetaLnx Live Demo



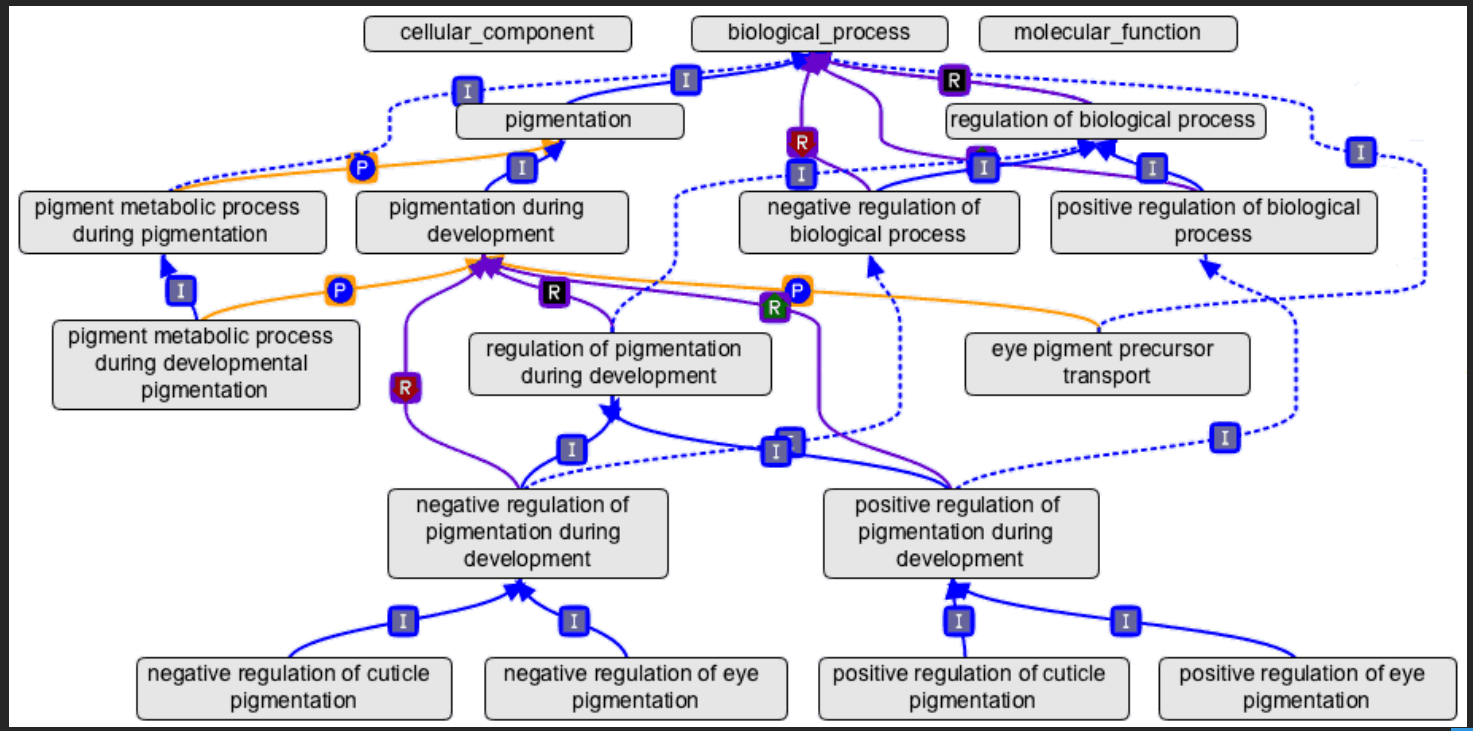
Metadata Hierarchy is Real


```
"Individual": {
      "id": "538764ae-5bc8-4f2b-b691-d99d6e1f754a",
      "groupIds": ["8c7409e2-d8be-4998-b00d-f7d5bee7c164"],
      "name": "John Smith",
      "description": null,
      "created": 1416416735,
      "updated": 1416416735,
      "species": {
         "ontologySource":
"http://purl.obolibrary.org/obo/NCBITaxon_9606",
         "id": "NCBITaxon:9606",
         "name": "Homo sapiens"
      },
      "sex": "MALE",
      "developmentalStage": {
         "ontologySource": "WHO:ICD",
         "id": "C18.5",
         "name": "Malignant neoplasm of splenic flexure"
      },
      "dateOfBirth": 1400324354,
      "diseases": [{
         "ontologySource": "WHO:ICD-O",
         "id": "8140/3",
         "name": "adenocarcinoma, NOS"
      }],
"phenotypes": [{
         "ontologySource": "ICD10",
         "id": "I12.9",
         "name": "hypertensive renal disease"
      }],
      "stagingSystem": "Dukes C",
      "clinicalTreatment": null,
      "strain": null,
      "info": {}
   },
"Experiment": {
      "id": "a93ba556-1e31-4d8c-bfe8-575a562a961c",
      "name": "Whole Genome Sequencing of Charge-S ARIC
sample A09994",
      "description": null,
      "created": 1416413182,
      "updated": 1416413182,
      "runDate": 1416413182,
      "molecule": "genomics DNA",
      "strategy": "whole genome sequencing",
      "selection": "random",
      "library": "IWG_IND-ARXREB.A09994_1pA",
      "libraryLayout": "Paired",
      "instrumentModel": "Illumina HiSeq 2000“
}
```


Partial example from an old GA4GH metadata use case

With complex associations...

Cross reference example from the the Gene Ontology Consortium



But changing….

- From a later GA4GH Metadata team report:
  - No single ontology works for all use cases
  - Complex nesting needed e.g. disease state and disease stage relationships
  - Multiple ontologies may be associated for aspects of a disease

- Recent work is shifting to a metadata framework which ties to other tools for ontology specifics

ON THE ABANDONMENT OF THE DEWEY DECIMAL SYSTEM

2012-11-15 · by Erin Lee Barsan · in Libraries & Librarians, Trends & Technologies.

Gone are the days of the card catalog. Is Dewey next? [Princeton University Archives]

While many public libraries stopped using the Dewey Decimal Classification (DDC) system to group fiction books long ago, the last five years have seen an emerging trend of libraries choosing to go "Dewey Free" for non-fiction books as well. In lieu of the industry standard system, these libraries are utilizing more so-called, "user-friendly" bookstore-style layouts. The 3,000 categories in this bookstore system model, Book Industry Standards and Communications (BISAC), are far less expansive and specific when compared to the 27,000 categories used to classify books with Dewey. The main features of BISAC include displaying books by their covers whenever possible, utilizing signage to spotlight current popular categories and grouping books by subject

Dewey Decimal Classification System

General Guidelines Below

000 Generalities	100 Philosophy and psychology	200 Religion
010 Bibliography 020 Library & information sciences 030 General encyclopedic works 040 Unassigned 050 General serials & their indexes 060 General organizations & museology 070 News media, journalism, publishing 080 General collections 090 Manuscripts & rare books	110 Metaphysics 120 Epistemology, causatio 130 Paranormal phenomen 140 Specific philosophical s 150 Psychology 160 Logic 170 Ethics (moral philosoph 180 Ancient, medieval, Orier 190 Modern Western philos	
300 Social sciences	400 La	
310 General statistics 320 Political science 330 Economics 340 Law 350 Public administration 360 Social services; associations 370 Education 380 Commerce, communications, transport 390 Customs, etiquette, folklore	410 Linguistics 420 English & Old English 430 Germanic languages G 440 Romance languages Fr 450 Italian, Romanian langu 460 Spanish & Portuguese 470 Italic languages, Latin 480 Hellenic languages, Cla 490 Other languages	
600 Technology (applied sciences)	700 Th	
610 Medical sciences and medicine 620 Engineering & allied operations 630 Agriculture 640 Home economics & family living 650 Management & auxiliary services 660 Chemical engineering 670 Manufacturing 680 Manufacture for specific uses 690 Buildings	710 Civic & landscape art 720 Architecture 730 Plastic arts, sculpture 740 Drawing & decorative ar 750 Painting & paintings (m 760 Graphic arts, printmakin 770 Photography & photogra 780 Music 790 Recreational & perform	

COMPUTERS

Use "**COMPUTERS** / Electronic Commerce" for works on the computer skills and technology needed to facilitate electronic commerce and use subjects beginning with "**BUSINESS & ECONOMICS** / E-Commerce" for works discussing the business aspects of electronic commerce.

COM000000 COMPUTERS / General

COM082000 COMPUTERS / Bioinformatics

COMPUTERS / Business Software *see* Enterprise Applications / Business Intelligence Tools

COM006000 COMPUTERS / Buyer's Guides

COM007000 COMPUTERS / CAD-CAM

COM008000 COMPUTERS / Calculators

COM009000 COMPUTERS / CD-DVD Technology

COM055000 COMPUTERS / Certification Guides / General

COM055010 COMPUTERS / Certification Guides / A+

COM055020 COMPUTERS / Certification Guides / MCSE

COM061000 COMPUTERS / Client-Server Computing

COM091000 COMPUTERS / Cloud Computing

COM010000 COMPUTERS / Compilers

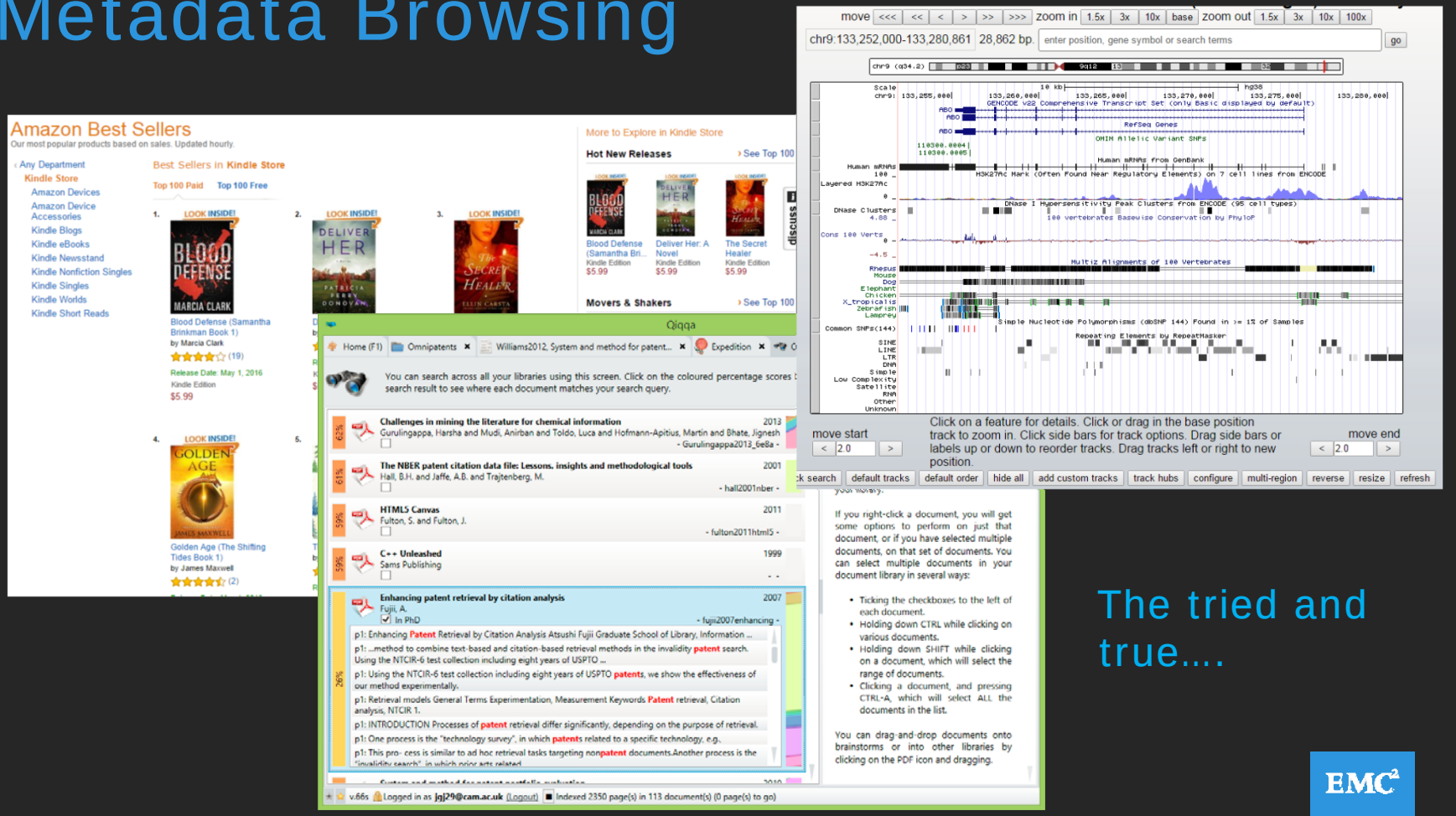
COM059000 COMPUTERS / Computer Engineering

COM012000 COMPUTERS / Computer Graphics

COMPUTERS / Computer Industry *see* BUSINESS & ECONOMICS / Industries / Computers & Information Technology

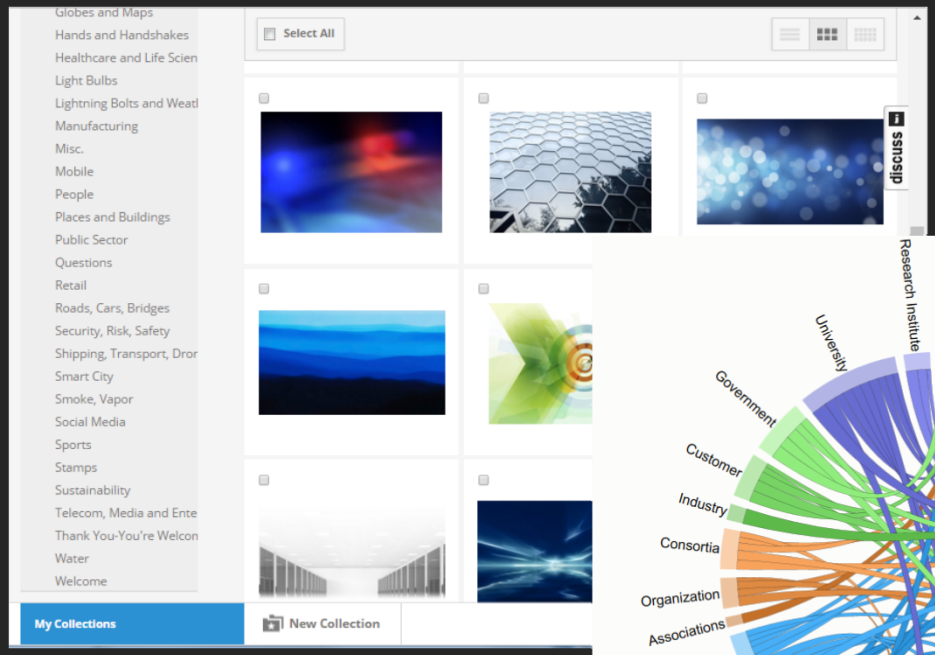
DDS which dates from 1876 is slowly being replaced by BISAC (Book Industry Standards And Communication)

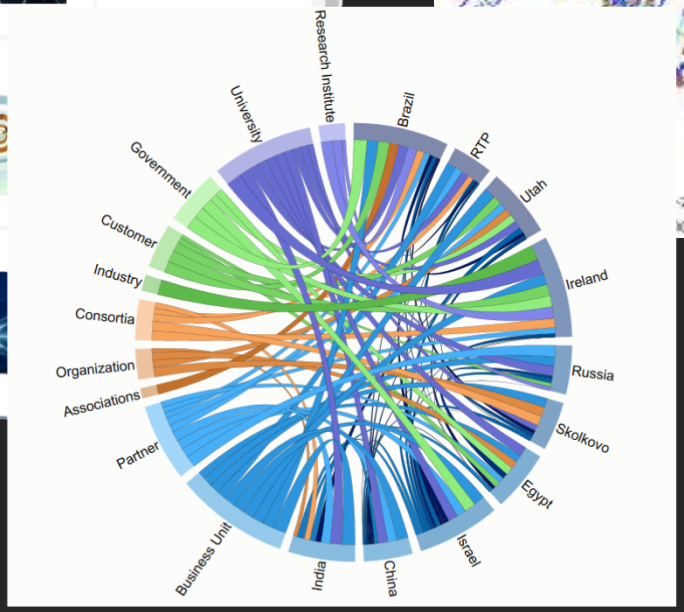
Metadata Browsing



The tried and true….

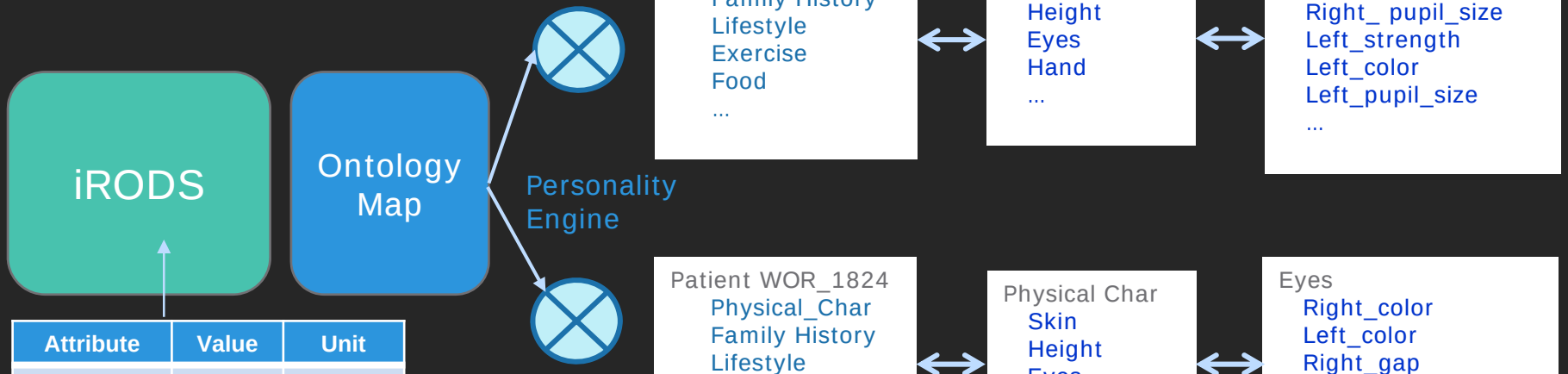
Metadata Browsing





Replaced by something new!

Tailored Metadata Browsing



Attribute	Value	Unit
Attribute 1	Value 1	
Attribute 2	Value 2	Unit 2
Attribute 3	Value 3	
...		...

EMC²

Questions