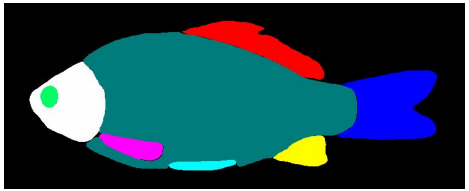
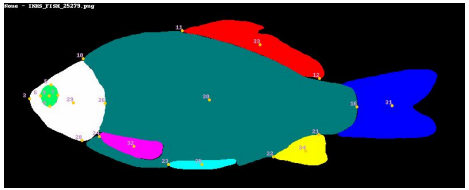
RDF Frameworks for Extensible Data Ingestion and Porting

Drexel University Metadata Research Center
Tulane University Biodiversity Research Institute
Dom Jebbia, Xiaojun Wang, Yasin Bakis, Henry L. Bart, Jane Greenberg
10/31/2022



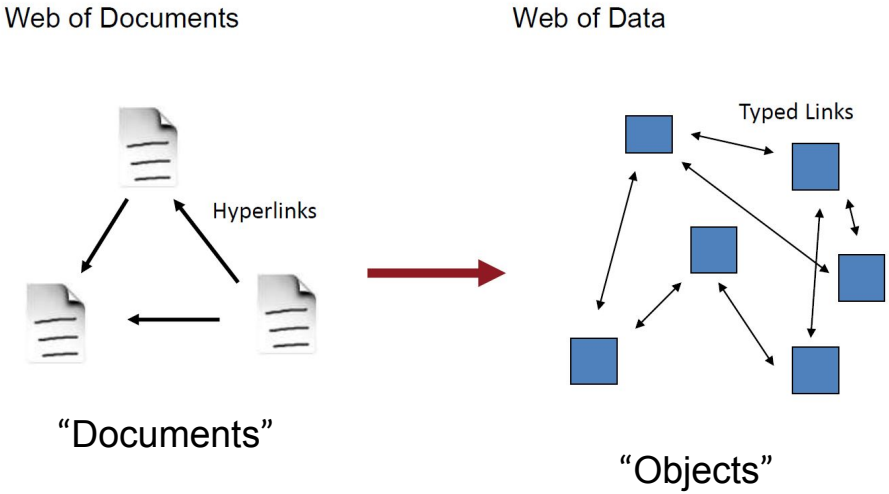


Raw image credit: Illinois Natural History Survey

Problem Space

- TUBRI has 300,000+ specimen images of fish*
- Metadata is derived from:
  - Collection event (specimen)
  - Raw image
  - Processed image
  - Labeled segmentation mask
- Numerous metadata standards
  - Dublin Core, Darwin Core, Audubon Core, XMP, EXIF, Photo Metadata Standard, etc.
- Need a flexible, extensible schema for database design

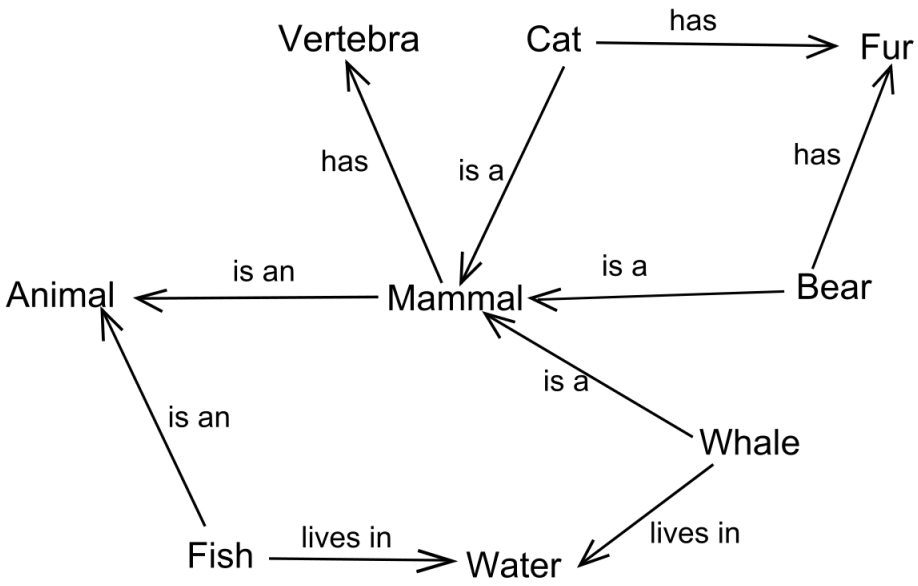
Linked Data



- Web of Documents
  - Names/URI
  - Documents described by XML, HTML, etc
  - Interactions via HTTP
  - (Hyper)linking between documents
- Web of Data
  - Names/IRI
  - Describes relationship between objects
  - Objects structured as RDF(Triples, Turtles, XML, JSON, etc.)
  - Linking and structure of data made explicit

Image credit: Fensel 2013

Semantic Network Graph



RDF Triples

<Fish> <is an> <Animal>.

<Fish> <lives in> <Water>.

<Whale> <lives in> <Water>

<Whale> <is a> <Mammal>.

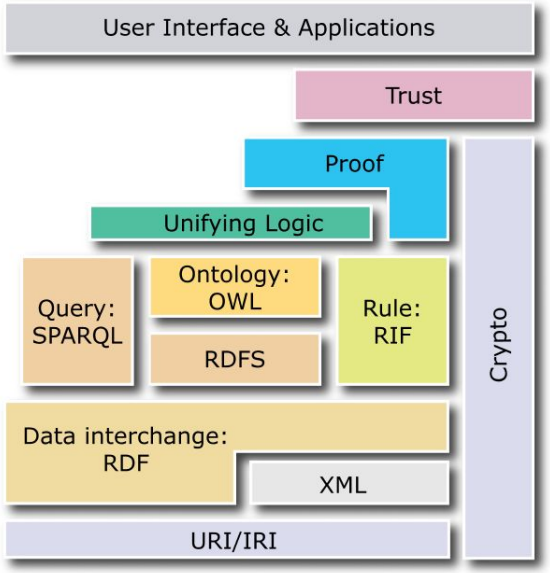
<Cat> <has> <Fur>.

<Bear> <has> <Fur>.

<Mammal> <has> <Vertebra>.

Image credit: Wikimedia

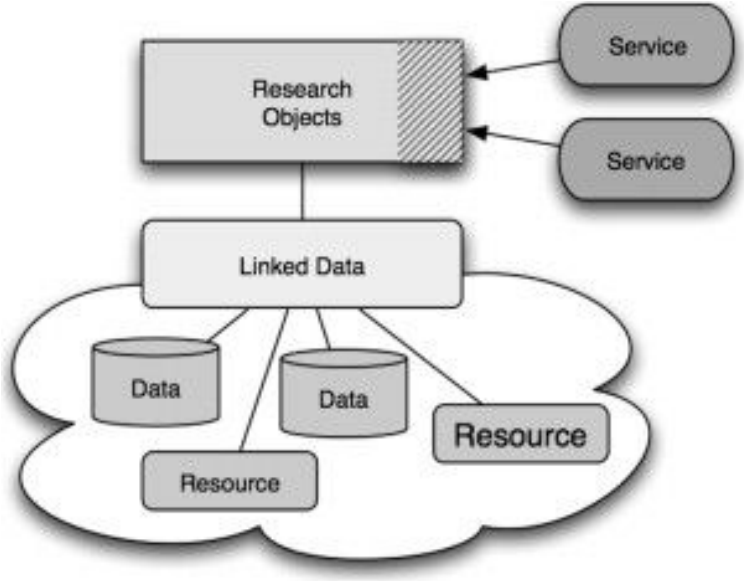
- Resource
- Description
- Framework



- A general framework for representing interconnected data on the web.
- Computer readable, NOT human readable.
- Creates a semantic graph network through linked data
- Flexible and easily extensible.
- DBpedia, Wikidata, FOAF, SKOS, etc.

Image credit: WC3

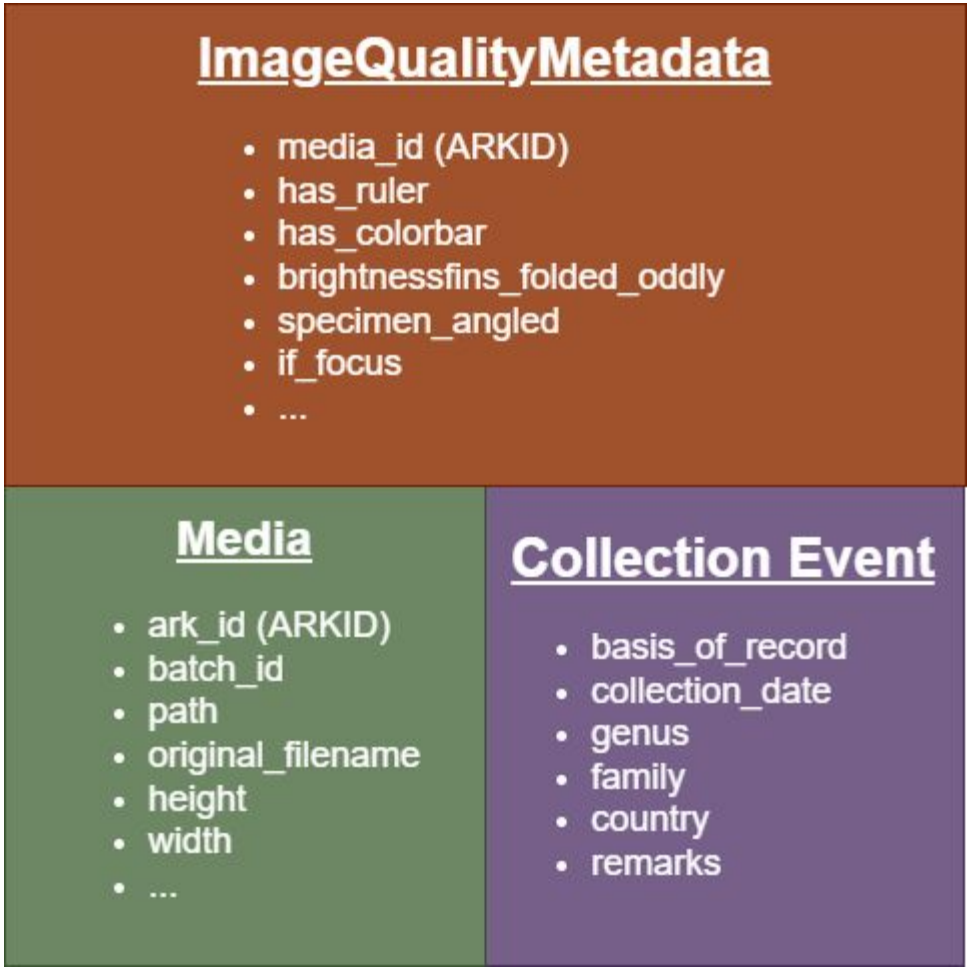
RDF Benefits



- Flexibility as data sources change
- Extensibility as data sources provide new (meta)data
- Quickly adapt to new technical challenges i.e. naming convention issue
- Links data to larger semantic and knowledge networks

Image credit: Beckhofer 2013

Original Structure



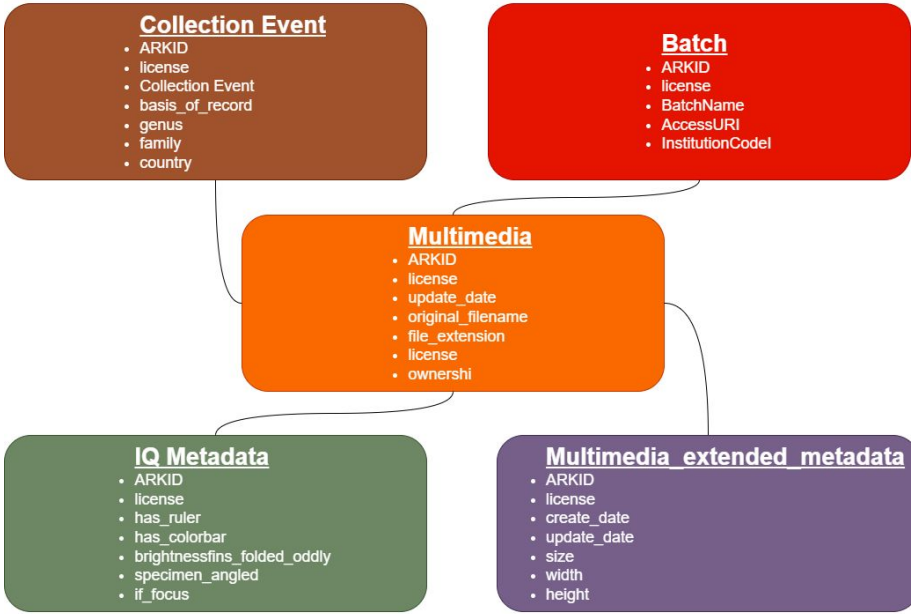
- Evaluated metadata elements in use.
- Examined other RDF implementations.
- Reviewed workflows for generating RDF.

- Xiaojun finalized Dom's recommendations
- Created prototype RDF/XML to build on.
- Used Protégé to generate RDF/XML .

Investigate → **Discover** → **Implement**

- More descriptive standards.
- Usually specialized; highest quality were living documents.
- RDFS in Python; Protégé used GUI + load from Excel.

RDF Prototype



- ~31 elements accompanied images (Collection event)
- TUBRI generated metadata based on processed images and bounding box images.
- Relationship between this metadata (objects) needs to be established.

What next?

- Refine protoype
- Look for better workflows
- Make RDF/XML available with data

RDF Frameworks for Extensible Data Ingestion and Porting

Dom Jebbia
djebbia@andrew.cmu.edu
Drexel University Metadata Research Center
Tulane University Biodiversity Research Institute
Carnegie Mellon University
10/31/2022

Supported by NSF-HDR-OAC: Biology-guided Neural Networks for Discovering Phenotypic Traits: 1940233 and 1940322m, NSF HDR-OAC:Imageomics: A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning: 2118240, and the Institute of Museum and Library Services (IMLS) RE-246450-OLS-20

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics