Identification and the Semantic Web

Greg Riccardi
Florida State University

Overview

- What and why of identification
- Familiar identifiers and information architecture
- Resolution and metadata
- Uniqueness
- Persistence
- Identifiers in collections management
- Semantic web and linked data
- Identifier resolution services

What and why of identification

- An identifier is a string that identifies a thing
- Some terms:
  - *Persistent*: an identifier can never be reused for a different thing
  - *Unique*: Within a community, the identifier is used by only 1 organization
- Identifiers allow information to be aggregated
  - Metadata from 2 sources about the same ID is about the same object
- Identifiers support services
  - Each community creates standards for which services will be available
- Start with some familiar examples

iDigBio
Integrated Digitized Biocollections

Familiar Identifiers

- Here we'll look at some examples of commercial identification strategies including UPC and DOI
- Universal Product Code
  - Bar code symbology
  - UPC is not globally unique, but rather a number that may be used by other systems for identification
  - UPC symbology has embedded redundancy that supports error detection
  - International standards organization (GS1) provides for allocation of identifiers.

Familiar Identifiers

iDigBio
Integrated Digitized Biocollections

Info. Arch. of UPC

- Example of use of UPC in retail
  - UPC can be used to access price for end user
  - UPC identifies product and not individual object

Info. Arch. of UPC

- Example of use of UPC in retail
  - UPC can be used to access price for end user
  - UPC identifies product and not individual object
- Inside the business
  - Package of 3 has contents
  - Manufacturing process is related
  - What other information is maintained and accessible via UPC
- Tracking of individual objects
  - In some cases, individual objects must be tracked
    - UPC is not enough and does not scale
  - Additional identifier is added, with its own bar code

Info. Arch. of UPC

- Example of use of UPC in retail
  - UPC can be u
  - UPC identifies ct
- Inside the busi
  - Package of 3
  - Manufacturing
  - What other inf essible via UPC
- Tracking of individual objects
  - In some cases, individual objects must be tracked
    - UPC is not enough and does not scale
  - Additional identifier is added, with its own bar code

DOI

- DOI, Digital Object Identifiers
  - Publication identification
  - ISO standard
  - "Semantic interoperability"
- DOI identifies various objects
  - doi:10.1038 is Nature publisher
  - doi:10.1038/ng0609-637 is "Genetics of reproductive lifespan," by Patricia Hartge in *Nature Genetics*, **v. 41**,
- Services are supported
  - Resolution via redirection (or "proxy")

    http://dx.doi.org/10.1038/ng0609-637
  - Standard metadata via LinkedData (simplified)

    <**title**>Genetics of reproductive lifespan</**title**>
    <**volume**>41</**volume**>

iDigBio
Integrated Digitized Biocollections

Resolution and metadata

- Getting information from UPC at retail outlet

iDigBio
Integrated Digitized Biocollections

Resolution and metadata

- Getting information from UPC at retail outlet
- Getting information from DOI on website
- Linked data for accessing metadata and objects
  - "A term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." Wikipedia
- Key technologies that support Linked Data are
  - URIs (a generic means to identify entities or concepts in the world),
  - HTTP (a simple yet universal mechanism for retrieving resources, or descriptions of resources), and
  - RDF (a generic graph-based data model with which to structure and link data that describes things in the world).

iDigBio
Integrated Digitized Biocollections

Exercise: Find some identifiers

- Go back to your online databases and find the identifiers of the various objects
  - Some identifiers are local (e.g. primary key)
  - Some identifiers are globally unique
  - Some identifiers are URIs
- List the identifiers and their characteristics (as above)

Uniqueness

- Uniqueness can be guaranteed
  - by context as in UPC, ISBN, DOI
  - by design: URI based on scheme plus DNS
  - By sparseness as in UUID
- Uniqueness can be reinforced by encoding
  - As in UPC, make values sparse
- Cannot reinforce single identifier per object

Persistence

- "Persistence" refers to the binding of identifier to object
  - Not object availability
  - An unexpected interpretation
- A persistent identifier is one that can be relied on for its connection to an object.
  - Once assigned to 1 object it will never be assigned to another

iDigBio
Integrated Digitized Biocollections

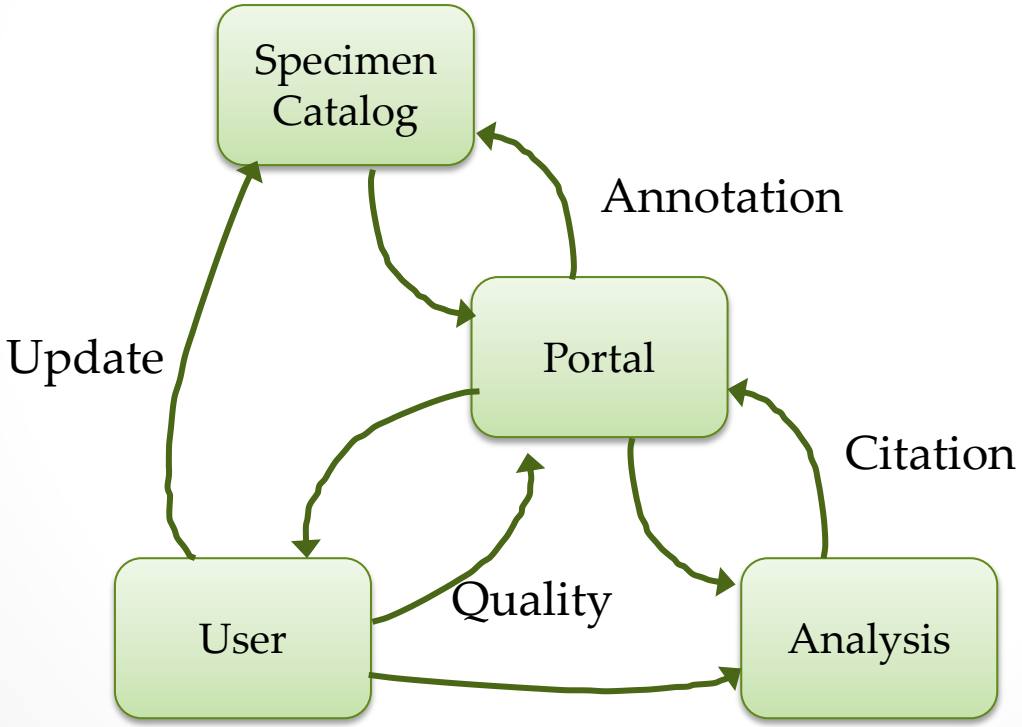
Annotation

- Persistent and unique identifications supports links
  - Comments on quality of object
  - Record of use of an object, e.g. in a publication

iDigBio
Integrated Digitized Biocollections

Info. Arch. for biodiversity

- Diagrams and discussion on the use of identifiers to track usage, quality control, and redundancy



iDigBio
Integrated Digitized Biocollections

Benefits of identification

- Data quality feedback
- Dialog based on annotation
- Tracking objects through analysis and use
- Maintaining attribution to provider
- Etc.

Identifiers in collections management

- GBIF uses Darwin Core Triple for determining uniqueness
  - (Institution, Collection, XX)
- GBIF now advocates adding a separate identifier to each occurrence record
  - What's wrong with Darwin Core Triple?

Choices of identifier

- Identifier should be a URI (universal resource identifier)
- Choose a URI scheme
  - http
  - lsid
  - doi (costs $$$)
- Big choice: embed information in identifier or not
  - UUID: assured uniqueness, completely opaque
  - Darwin Core triple: http://nybg.org/herbarium/1123324
  - Combination?
- Advantage of embedding information
  - Easier to provide resolution service
- Disadvantage
  - People infer meaning from what they read

Semantic web and linked data

- Linked Data allows agents to find and aggregate information
  - “Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF."
- Initial implementation based on RDF (more later), HTTP URIs and HTTP protocol
- An HTTP get requests mime-type via accept parameter
  - Accept: application/rdf+xml

iDigBio
Integrated Digitized Biocollections

Problems with Identification

- Semantic web allows agents to infer facts from identified information
- Sometimes this is very hard and context dependent.
- For example, ISBN identifies a "book"
  - Different versions have different ISBNs.
  - If I am interested in reading the book, I don't care. Each ISBN for the same "book" identifies an equivalent object
  - If I am a book collector, I might want a first edition, or the hardback, or one signed by the author.
- Examples from collections?

Summary/conclusions

- Information architecture must be planned.
- Identification is key to information quality management.
- Identification allows tracking of citations and updates.
- Information services depend on identification.