Completeness, Coverage & Equivalence in Scientific Data Records

Andrea K. Thomer, Karen S. Baker, Simone Sacchi & David Dubin
{thomer2, ksbaker2, sacchi1, ddubin} @illinois.edu
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St. Champaign, IL 61820-6211 USA

ABSTRACT
Previously we asked, "When is a record data and when is it a fish?" (Wickett et al., 2012). In this work, we ask, "when and in what contexts are a record and a fish equivalent?" We describe and compare a collection of potentially equivalent records describing a *Mola mola*, or Ocean Sunfish, specimen. We calculate the Metadata Coverage Index (MCI) of each record and explore the use the Systematic Assertion Model (Dubin, 2010) to support investigation of the assertions contained in these data records.

Keywords
Metadata coverage, biodiversity informatics, scientific equivalence

INTRODUCTION
Natural history museum specimen records are increasingly provisioned and discovered online through cloud-hosted databases such as GBIF and VertNet. While increased use of standard vocabularies like Darwin Core means that these records are more easily aggregated and made interoperable (Wieczorek et al., 2012), the act of cross-walking legacy data and then transferring records from local to cloud-based databases with different representation formats, encodings, and harvesting protocols results in the proliferation of different versions of the "same" record. Depending on the vocabulary and/or schema used, these roughly equivalent records make different amounts and types of data available, and, thus, their fitness-for-purpose or analytic potential in different contexts varies (Hill et al., 2010; Palmer, Weber & Cragin, 2011).

In prior work (Wickett et al., 2012) we considered a *Mola mola* species occurrence record pulled from a Darwin Core Archive (DwC-A) file available on the University of Kansas Biodiversity Institute Integrated Publishing Toolkit (KUBI IPT) installation to explore these issues ("Gbif.org - Darwin Core Archives"). Here, we compare five records downloaded from different data providers describing this same specimen, explore the metadata coverage and completeness of these records, and more fully discuss the nuances of determining their equivalence. We explore the use of the Systematic Assertion Model (SAM; Dubin, 2010) to compare conflicting assertions between data sources when they arose.

The Dataset
Our original record *Mola mola* record was chosen from the DwC-A because sunfish are a “charismatic megafauna” with which we were somewhat familiar. We found additional versions of this specimen record through different data providers: another locally hosted by KUBI consisting of a "web view of the local database record... before the crosswalk" to DwC (L. Russell, pers. comm., 26 April 2011), and others in GBIF, FishNet and VertNet (Figure 1).

ANALYTICAL APPROACH
After qualitatively assessing each record, the content of all five records was aggregated into a spreadsheet to better provide an overview of their similarities and discrepancies; the Metadata Coverage Index (MCI), or "the total number of 'non-missing' fields expressed as a percentage of the total fields available across all records" (Liolios et al., 2012) was calculated for each record (Table 1). Both a “composite” and "native" MCI were calculated, the native MCI being the total number of non-missing fields expressed as a percentage of total fields available across the schema of the individual record’s native database.

Scans of the original field notes (written by the specimen’s collector, Martin Wiley) describing this specimen were located, along with photographs; the field notes and physical specimen represent the source of the data record. Because the field notes are formatted as free text journal entries and species lists, they were not aggregated into the
spreadsheet with the other records, but instead, were qualitatively compared. Finally, we consulted with University of Kansas Ichthyology (KUI) collection manager Andy Bentley who catalogued the specimen in 2003, when it and a number of other specimens were unexpectedly discovered in the collections.

<table>
<thead>
<tr>
<th>Source</th>
<th>Total Fields</th>
<th>Complete Fields</th>
<th>Composite MCI</th>
<th>Native MCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>KUBI IPT</td>
<td>41</td>
<td>32</td>
<td>55%</td>
<td>78%</td>
</tr>
<tr>
<td>KUBI Web</td>
<td>18</td>
<td>15</td>
<td>37%</td>
<td>83%</td>
</tr>
<tr>
<td>GBIF</td>
<td>25</td>
<td>22</td>
<td>25%</td>
<td>88%</td>
</tr>
<tr>
<td>VertNet</td>
<td>22</td>
<td>10</td>
<td>22%</td>
<td>45%</td>
</tr>
<tr>
<td>Fishnet</td>
<td>28</td>
<td>13</td>
<td>17%</td>
<td>46%</td>
</tr>
</tbody>
</table>

Table 1. Comparison of the Composite and Native MCI’s of each record. The “Composite” metric is out of a total 58 aggregated fields.

PRELIMINARY RESULTS
The record from KUBI’s IPT installation seemed initially most informative, with 32 of 41 available fields filled out and a native MCI of 78%. Though this record did not have the highest native MCI (GBIF’s did, at 88%), it did have the highest composite MCI (55% versus GBIF’s 25%). However it did include one potentially problematic data field: the record’s "basisOfRecord" is a "PreservedSpecimen." Remember, the Mola mola is a megafauna: adult sunfish can grow up to 4 meters in length and weigh up to two tons (“Ocean Sunfish - EOL”). While it’s conceivable that a full-grown specimen could be preserved in a museum collection, it would be unusual.

KUBI's locally hosted record (the “web view” of the collections database) made sense of this impossible-seeming specimen; it describes the specimen's “size” as "larva." Photographs sent by the KUI collection manager confirmed that this was indeed the case -- a very young Mola of approximately 1 cm in length (Figure 2). This closer re-examination led him to question Wiley’s 1965 identification of this specimen. Fish larvae of this family and at this age look quite similar to one another; consultation with an expert on Molidae (which Wiley was not) may be necessary to determine the fish’s taxonomy with greater certainty. The locally hosted KUBI record was the only one that contained data about the specimen's size or age.

DISCUSSION AND ANALYSIS
The necessity of, and potential for problems caused by, the missing data describing the specimen’s size and age became apparent when we learned from Bentley that he had reason to believe this Mola may not even be a Mola mola. The uncertainty in Wiley’s identification was almost entirely due to lack of documentation about its young age; adult Mola mola are far more distinct than the young (ibid). Even if this specimen has been

Figure 1. Chronology Mola mola of record proliferation.
correctly identified, missing data about growth phase can greatly impact later analysis. Just as a caterpillar inhabits a very different environment than a butterfly, a juvenile Mola could inhabit a different niche than an adult. Thus, the absence of age data may mean that this occurrence record is unfit for use in some scientific investigations.

The absence of the age field in the GBIF record, at least, is likely attributed to GBIF’s metadata provisioning policies; due to processing constraints, they only display 25 fields of data, and “dwc:age” is not one of them (T. Robertson, pers. comm., 29 Mar 2012). The FishNet and VertNet records’ limitations could potentially be artifacts of GBIF’s policies, though this is uncertain.

**Completeness vs. coverage**

Even a record with an MCI of 100% within its native database could nevertheless be missing vital data; the GBIF record contains neither an “age” field nor extensive georeferencing data, yet has the highest native MCI of our recordset. Thus, we must be careful not to confuse a record’s metadata coverage index with its completeness; additionally, a high MCI is not a suitable metric in determining record’s fitness for purpose. Though the MCI seems promising as a simple metric of quality, additional, more sophisticated, methods of analysis are needed to account for a dataset’s strengths and limitations in the context of other datasets.

The Systematic Assertion Model (SAM) may offer us one such method by providing a logic-based framework to reason over the scientific claims expressed by a dataset. In this particular case we applied SAM to compare the two...
conflicting assertions concerning the Mola’s taxonomy by VertNet (appealing to Wiley’s field notes) and by the collection manager who re-assessed Wiley’s original data (Figure 3).

FUTURE WORK
We intend to apply this methodology to a wider range of specimen records to further explore the data transformations and losses that occur when specimen records are crosswalked from one format and/or vocabulary to another. We will continue to explore the use of SAM as a way to augment simple metrics like the MCI; the ability to determine scientific equivalence as well as metadata coverage could be integral to the development of repository or aggregators’ dashboards (e.g. Fenlon et al., 2011).

Additionally, it’s worth noting that KUBI’s IPT installation is neither a formal nor public data source; it’s a staging ground for provisioning records to GBIF and other aggregators, and thus a unique middle ground between local and remote repositories or data providers. Yet this unofficial, obscure data source is the most informative (despite the lack of one crucial field of data). We note that the reality of practice in data provisioning dispels the notion of a well-defined linear path from data origin to data publication -- one that merits further investigation of the kinds of interdependencies identified by Baker and Yarmey (2009) as a “web of repositories.”

ACKNOWLEDGMENTS
Thanks to Karen Wickett and Nic Weber for helpful feedback in the preparation of this manuscript. Thanks also to Andrew Bentley and Laura Russell at the University of Kansas for providing us with photographs, field notes, and conversation.

REFERENCES