Flags are not enough: A friendly spreadsheet user interface for Quality Control of occurrence data.

David B. Lowery[1], Tracy Barbaro[2], James Hanken[3], Sven Koehler[4], Bertram Ludaescher[5], James A. Macklin[6], Paul J. Morris[7], Robert A. Morris[8], Laura Russell[9], Tianhong Song[10], and John Wieczorek[11]

[1,3]Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA;
[2]Learning and Education Group, Encyclopedia of Life Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA;
[4]Dept. of Computer Science, University of California, Davis, USA;
[5]University of Illinois Urbana-Champaign, Champaign, USA;
[6]Agriculture and Agri-Food Canada, Ottawa, Canada;
[7]Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA;
[8]Dept. of Computer Science, University of Massachusetts at Boston, Boston, MA, USA;
Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA
[9]Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark;
[10]Dept. of Computer Science, University of California, Davis, USA;
[11]Museum of Vertebrate Zoology, University of California, Berkeley, USA;
Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA

Abstract

Scientists who use a large collection of natural science data sets may be satisfied with data quality reports that allow inclusion or exclusion of subsets of the data from particular use. Choices may be based on one or more criteria supporting the fitness for the user's purpose. By contrast, a data curator will be much more likely to need information concerning the provenance of data quality assertions in order to rationally act on those assertions. We describe a friendly spreadsheet-based data quality report produced by tools developed by the FilteredPush and Kurator projects. The structure of the report is influenced by the needs of data curators and data users. The report is in the form of a multi-page spreadsheet containing a summary sheet and sheets reporting the details of data quality for particular conceptual operations on the data (e.g. georeference validation). The detailed report sheets include flags to assert the status of each row of data, original data and proposed corrections. Also included are a list of external sources that were consulted in the evaluation of each row and a human readable comment that reflects a provenance trace through the internal logical steps taken by the data quality software.

1 Introduction

Data about occurrences of organisms at particular places and times support many scientific endeavors ranging from environmental studies to evolutionary hypotheses to climate change studies. Both users and managers of such data have a distinct need to understand the fitness of such data for the use to which

they will put it. These needs arise both about specimens preserved in natural science collections as vouchers for the occurrence, or unvouchered human or machine observations. Users and data managers are also concerned with the ease, reliability, and repeatability with which that data can be accessed and aggregated with related data. These concerns can all be discussed under the rubric of data Quality Control (QC), a discipline historically focused on manufacturing, which arguably originated with the 1922 publication by Radford of "The Control of Quality in Manufacturing" [Radford, 1922]. Radford's preface relates:

> In the factory, quality is a costly thing to neglect, yet it is the usual experience to find a disproportionate emphasis placed upon quantity of output, in the effort to effect economies. Often, this is not so much due to lack of proper intent as it is to the failure to realize what the quality approach means. [...] if quality is under positive and continuous control, increase of output follows as a by-product advantage.

In the first few decades of natural science collection digitization the emphasis was on careful, rich transcription of all of the details of paper records associated with specimens. In the 1990s, recognition arose that specimens were entering collections at a faster rate than digital records could be created with such detail and high data quality. Hence, particularly in recent large scale digitization (e.g. [iDigBio, 2011] and [Haston et al., 2012]), focus has shifted to rapid transcription of fewer data elements, with an emphasis on making dark data accessible [Smith et al., 2012, Smith and Blagoderov, 2012].

2 Shifting historical foci of concern

As Chapman pointed out over 10 years ago [Chapman, 2005], Quality Control of biotic occurrence data shares many QC concerns with those of other science disciplines and those of manufacturing. In the context mentioned above, we could recast Radford's preface as:

> In the natural science collection, data quality is a costly thing to neglect, yet it is the usual experience to find a disproportionate emphasis placed upon quantity of output, in the effort to effect economies.

This formulation evokes Shewhart's Plan-Do-Check-Act model [Shewhart, 1939], wherein a QC application performs checks and the resulting outcome is a basis for Act. A more modern, explicitly *data* quality control model, would phrase this as the Define-Measure-Analyze-Improve meme in [Wang, 1998].

Albeit with some resistence [Shetler, 1974], natural science collections digitization began in earnest in the 1970s (e.g. with SELGEM [Creighton and Crockett, 1971]; ** TODO: CITE NSF collections computerization reports **), with a prescient vision (** TODO: CITE, malacology standard **) of the power for research of data exchanged and aggregated across all collections. In the 1980s, there was widespread recognition that this long term vision of data interchange needed common understanding of concepts and vocabularies across collections and documents. Influential models for common concepts emerged in the late 1980s and early 1990s, in particular the ASC model [Assn. of Systematics Collections, 1992], and early versions of HISPID [HISCOM, 2013] and ABCD [ABCD, 2015]. In 1985, the Taxonomic Databases Working Group (TDWG) emerged as a forum for global standardization efforts in the domain [TDWG, 2007]. (Although the acronym remains in wide use, the organization is a body now named Biodiversity Information Standards (TDWG)). An emphasis of TDWG in the 1980s and 1990s was the development of controlled vocabularies to promote the exchange and integration of taxonomic data.

Within the TDWG Data Quality Interest Group (BDQ) BDQ, a conceptual framework for describing data quality needs, solutions, and reports has been developed. The framework description is described in Veiga

et al. [2017] This framework (FFDQ herein) derives from Wang's approach and provides a formal structure for describing data quality needs, data quality mechanisms, and data quality reports. The present work relates to the data quality reports of the framework, and can be presented in terms of the framework Kurator-FFDQ and API Kurator-FFDQAPI.

Somewhat inaccurately, we could characterize this phase of the development to the digitization of natural science collections data as Plan-Do. That is, *plan* the designs for information systems and controlled vocabularies for data capture, and *do*, build them and populate them with data. An emphasis in the 1990s was to design information systems that limited the ability of data entry personnel to make mistakes [Morris, 2005]. Networks developed first for distributed query (e.g. FishGopher [see Wiley and Peterson, 2004, p.92], MaNIS [Stein and Wieczorek, 2004, Wieczorek, 2015], VertNet [VertNet, 2015], HerpNET [HerpNET final report], FishNet [FishNet web site], ORNIS [ORNIS web site][1], BioCase [BioCASE, 2015]). These were followed quickly by networks and other tools for data aggregation of natural science collections data (OBIS [OBIS, 2011], GBIF [GBIF, 2015], iDigBio [iDigBio, 2015]), along with shifts in the emphasis in digitization strategies from rich data capture to rapid data capture from bulk data sources (e.g. transcription from handwritten ledgers). Consequent data growth and availability led to an increasing recognition of data quality issues (e.g. [Frey, 2006]). At the same time introduction of networks of unvouchered observational data (e.g. eBird [Sullivan et al., 2009] and [Ahumada et al., 2011]) led to even more rapid data growth. With these devlopments has come the understanding that fitness for reuse demands more and more emphasis on data quality control (e.g. [Chapman, 2005]). Along with classical taxonomic uses, species occurrence data is valuable for research questions related to global change, ecological studies, and agricultural applications (e.g. [Brooke, 2000, Graham et al., 2004].) In the face of these changes, the community has shifted further towards Check-Act, *checking* data quality and *acting* to improve it (e.g. [CLO, 2012].)

TDWG collaborates with the Global Biodiversity Information Facility (GBIF), an international treaty-based organization that provides an open data infrastructure funded by signatory governments. At this writing, GBIF's data portal aggregates almost 900 million occurrence records of over 2.5 million species. These records are aggregated and indexed over 37,235 datasets published by 1,130 data publishers worldwide. GBIF produces software tools and components, notably several for producing and managing Darwin Core Archives (see [GBIF, 2013]), that are capable of publishing occurrence data coded to TDWG's Darwin Core (DwC) controlled vocabulary [Wieczorek et al., 2012, TDWG, 2015]. These document both specimens and unvouchered observations. The latter are growing faster in number. For example, as of October 11, 2016, GBIF serves nearly four times as many observation records as specimen records, compared to a factor of three a year earlier.

Chronicaling the history of computerization at the Berkeley Museum of Vertrate Zoology, Sunderland [Sunderland, 2013] relates that institutions had already noticed that the costs associated with mainframe computing didn't justify systems devoted only to cataloging specimens. Computers would also have to support the storage and discovery of scientific knowledge as well.

3 Assertions vs "facts"

For over a century, practicing taxonomists have agreed that classification and description of biotic occurrences often involves professional opinions that are subject to revision and dispute:

> "There can be no precise criterion of species. 'Species' is a human concept as much as 'genus,' and of the same sort." ([Lutz, 1908] p.7.)

[1]The data portals of ORNIS, MaNIS, FishNet, and HerpNET have recently been merged with that of VertNet, although the underlying collaborations generally remain intact.

For this and other reasons, many artifacts of taxonomic data Quality Control are themselves represented as assertions about fitness for use, not necessarily affirmation or denial of factual correctness. Simple cases are the logical correlation between several data attributes sometimes assembled by the FilteredPush (FP) [FP web site] project's workflows. For example, in cases where we have access to services giving biographical data of collectors (e.g., the machine-accessible service Harvard Index of Botanists [Harvard University Herbaria, 2013], the workflow components may query such services to examine the lifespan of the names of the collectors named in the "dwc:recordedBy" attribute. If the collection date falls outside those lifespans, the workflow can only assert that the biographical and collection data cannot simultaneously be correct. This leaves it to the workflow user to decide which, if either, should be deemed correct and what to do about it. More technical details of the FP workflow components are to be found in [Song et al., 2014].

Throughout the paper, following common conventions, if a vocabulary term is defined in DwC we will preface it with "dwc:".

In this paper, we are principally concerned with a friendly spreadsheet-based report format aimed at helping data curators and other users evaluate the fitness for use of FilteredPush QC assertions generated by the Java application *FP-Akka* FilteredPushWiki, and related assertions produced by the Kurator-Akka application [McPhillips et al.] .

4 FP-Akka Architecture

Figure 1 describes the architecture of FP-Akka. The platform takes its name from its use of the open source Akka platform for high performance, asynchronous, distributed applications [Typesafe Inc, 2015]. Currently, two artifacts are provided for users to run. The first, *FP-Akka-workflowstarter jar*, is a Java application that takes various flat DarwinCore inputs, and creates intermediate JSON (JavaScript Object Notation) [ECMA International, 2013] output. The second, the *FP-Akka postprocessor* [Kurator Project, 2015] creates human readable spreadsheets from the intermediate JSON. The FP-Akka-workflowstarter application itself includes the *FP-Akka* component that describes a data flow workflow paralellized with the Akka framework, and the FP-KurationServices component, which contains the internal implementations of domain specific data validation actors. The latter invokes a range of external services provided by authoritative sources in the community. The code base of FP-Akka seeks to separate concerns of dataflow management, domain data quality business logic, web service invocation, and presentation to end users. The FP-KurationServices code had its origins in the Kepler-Kuration module [Dou et al., 2012, 2011], which presented a proof of concept of a data quality control workflow written within the Kepler workflow environment. Key elements of the metadata concerning data quality assertions by the code, the Curation

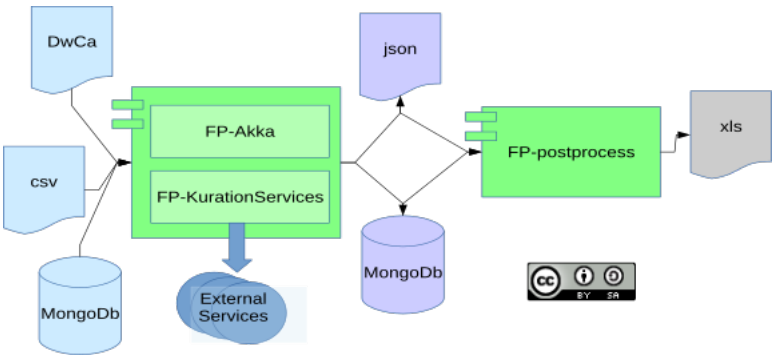

Figure 1: FP-Akka Architecture

Status, the Curation Comment, and the Services Consulted, have been retained from the Kepler-Kuration work. These represent an interface between the internal business logic of a data quality control actor, and the flow of data and metadata in the workflow. They also represent the key elements we have identified for clustering and interpretation of data quality assertions by end users. The curation status is the key sorting element for analytical uses (i.e., does this record meet inclusion/exclusion criteria based on a test of fitness for some purpose). The services consulted and the curation comment provide critical provenance information about each data quality assertion that allows a data curator to assess what actions, if any, to take on the data elements involved in that assertion. Since the time of the Kepler-Kuration implementation, a number of components have changed. These include the workflow framework itself, particular workflows and presentation mechanisms. Also, the internal quality control logic of the actors has been reworked. However, the *interface* for presentation of the assertions made by the quality control code to downstream consumers (e.g plugins to Symbiota and other specimen management systems, as well as the new postprocessor that is the subject of this paper) has remained unchanged. That interface remains described by the aforementioned intermediate JSON.

5 Validators

FP-Akka presently supports five principal data quality actors which are put together with input and output components for the purpose of deriving QC assertions about taxon occurrence data. These actors, which we call "validators," are the ScientificNameValidator, the EventDateValidator, the DateValidator [2] the GeoReferenceValidator, and the BasisOfRecordValidator. In terms of the framework FFDQ, these actors assert a mixture of measures, validations, and enhancements. The validators act on appropriately coded occurrence data, e.g. Darwin Core Archives ([Robertson et al., 2015]) or simple comma separated values (CSV [IETF, 2005])-coded text files. Most of the FP-Akka components assume that the input occurrence data are in the form of flat DarwinCore and are coded with attributes named, except for case sensitivity, by the Darwin Core term names. In Kurator-Akka, we have developed workflows, derived from the VertNet migrator tools, for aiding in the mapping of local data storage terms onto Darwin Core terms (Darwin Cloud), but in FP-Akka, this mapping was left as an exercise for the data provider. For example, the ScientificNameValidator can make QC assertions about dwc:scientificName and about the authorship of a name (dwc:scientificNameAuthor). The actors in FP-Akka consist almost entirely of code that manages the flow of data and data quality assertions through the workflow. All of the business logic of data quality control is separated into classes in the FP-KurationServices module. The FP-Akka actors are composed into workflows. Three workflows are in current production use in FP-Akka. The first is the standalone *DwCa workflow* that takes DarwinCore data in the form of CSV files or DarwinCore archives and passes it through the validator, and then writes the intermediate JSON output to a file.

The second DwCa workflow is illustrated in Section 9. It is intended to allow users to run analyses on their own data sets. A similar workflow is embedded in FilteredPush network nodes as an analytical capability (NEED A GOOD REF). It is identical to the standalone DwCa workflow except for data loading and writing actors that interact with MongoDB [MongoDB, 2015] instead of the filesystem.

The third workflow consists of a CSV reader, the ScientificNameValidator and a CSV writer. It is intended for checking taxonomic authority files in collection databases. This workflow takes a CSV dump from a natural science collection database taxonomic authority file comprising the primary key from the table, the scientific name, and the scientific name authorship. The workflow output retains the primary key value, and adds values for corrected scientific name and scientific name authorship if provided by a name authority. If the name authority offers a globally unique id (GUID) for *its* record, that GUID is added to

[2] See, e.g. http://rs.tdwg.org/dwc/terms/#Event for date forms available to the EventDateValidator and http://rs.tdwg.org/dwc/terms/#Taxon for some data available to the ScientificNameValidator.

the output as provenance. Also in the output are the original values for the scientific name and scientific name authorship, as well as the curation status. Added provenance called "Comment" is provided in the FP-Akka JSON output but labeled as "Provenance" in the spreadsheet models, as illustrated in Table 2. Each addition is rendered as a column in the output CSV file. This format provides a very effective tool for database managers, since a text pattern matching utility such as *grep* [Wikipedia, 2015] can isolate lines that contain a particular QC assertion. Each line contains the input value of the database primary key along with an assertion of a GUID found for the matched name in the specified authority. Hence, regular expression tools can readily convert a set of lines into SQL update statements that can be fired against the database to add the GUID to validated records (identfied by their database primary key). For example, the output line:

```
‘‘62740’’,‘‘Scientific Name Authorship was empty’’,‘‘changed to:  Walckenaer,
1805’’
```

from Table 2 is easily transformed into the SQL query

```
‘‘update taxon set SciName = ‘Latrodectus Walckenaer, 1805’ where taxonid = 62740’’.
```

Likewise grep can identify particular kinds of problems found in the dataset and the corresponding rows can be put into a spreadsheet and sent, as a trackable unit of work, to collection management staff for evaluation and correction of the relevant database records.

Some Darwin Core attributes have both an atomic form comprising several components and a federated form conforming to some data standard or authority. For example a record set may have values for the year, month and day of the collection event, and also a value for dwc:eventDate comprising a date string compliant to the DwC-recommended ISO 8601 date standard [ISO, 2004]. The FP-Akka DateValidator can compare an ISO 8601 date constructed from the three atomic fields to that given by (or missing from) dwc:eventDate.

Each validator produces its assertions as one of five outcomes along with provenance information, i.e., information about how and why the validator came to make those assertions. Of particular provenance interest is the authority for the outcome, including well-known authority servers and caches that a particular configuration may support. For example, a user who knows that the input comprises entirely plant data can arrange to consult specific plant-centric taxonomic name authorities such as the International Plant Names Index (IPNI) [IPNI, 2012]. Specialist knowlege of scope of the external resources consulted is important. IPNI is an authority on the nomenclature of vascular plants, but currently does not have strong coverage of non-vascular plants, or fungi (which are also covered by the botanical code of nomenclature, ICNafp [Wiersema et al., 2012]. Under the current architecture of FP-Akka, a holder of a botanical data set would have to make an informed choice about selection of IPNI or IndexFungorum (IF) [IF, 2015] based on their knowlege of the taxonomic scope of their data.

6 Outcome Values

Validation Outcome	Assertion about object validated	Workflow outcome label	Spreadsheet cell color
CORRECT	Nothing found wrong	no change needed; looks good to us	Green
CURATED	Proposed change	we have proposed this change	Yellow
FILLED_IN	Proposal for a missing attribute	no value present; we have proposed one	Mustard
UNABLE_DETERMINE_VALIDITY	Can't tell whether value is valid (generally, preconditions for validation were not met).	don't know	Gray
UNABLE_CURATE	Value is not validated but workflow can't suggest a change	there seems to be a problem, but we don't know how to solve it	Red

Table 1: FP-Akka Outcomes

The data quality control logic in FP-KurationServices classes (Figure 1) returns a validation state (for the evaluation of a particular set of fields for a particular record) with one of five values. These are passed on to the intermediate JSON as one of five strings ("Validation Outcome" column in Table 1). These values are "CORRECT", "FILLED_IN", "CURATED", "UNABLE_DETERMINE_VALIDITY", and "UNABLE_CURATE."

FFDQ expands upon this set to provide a richer set of outcomes classified by assertion type (validation, measure, amendment). Additionally, since the framework is focused more on the individual assertions rather than the actor logic as a whole, monolithic actors can be broken down into a series of assertions or tests, the dependencies between these tests, and the test outcomes.

In FFDQ the value "CORRECT" corresponds to a result, rather than a status, of "COMPLIANT" or "NON_COMPLIANT" for a Validation, "FILLED_IN" and "CURATED" are associated with Amendments. Two additional outcomes are included for Measures, "COMPLETE" or "INCOMPLETE" and the "UNABLE_DETERMINE_VALIDITY" and "UNABLE_CURATE" are treated as the assertion run status along with "NOT_RUN" and "AMBIGUOUS." This expands upon the original set of outcomes by defining a more structured run result with an associated status.

An assertion that a value is correct is a claim that the data quality control logic found no inconsistency between the data presented and any oracles that it consulted. This may mean that the data was validated, as in the case of the ScientificNameValidator finding a single exact match for a scientific name string and authorship in IPNI and being able to assert IPNI's LSID for the matching IPNI nomenclatural record. CORRECT may also mean that the data was merely consistent, as in the case of the collecting event DateValidator asserting that a particular collection date in dwc:eventDate falls within the lifespan of the collector in dwc:recordedBy as asserted by a biographical record for a person who's name matches the string found in dwc:recordedBy. The collecting event date could still be off by days or months or years, but the data quality code was not able to detect a problem.

An assertion that the outcome was FILLED_IN represents an orthogonal concern that we have not untangled in the FP-Akka development (but are addressing in ongoing Kurator project development [Kurator Wiki]). That is, a dwc:eventDate may have been empty, but a dwc:day, dwc:month, and dwc:year may have allowed us to construct an event date and compare it with any known collection dates for the collector. An assertion about the comparison of the dwc:eventDate with the collector's lifespan is an orthogonal concern to our having filled in dwc:eventDate from related data elements. This orthogonality highlights a defect in FP-Akka in that it can only present a single outcome for a given data record.

With FFDQ we can model this as two assertions, a Validation with a result of NON_COMPLIANT that leads to an Amendment with result FILLED_IN. Another added benefit is the ability to identify results at different stages of QC. Using the previous example, the additional information that could be included would be the state of the validation prior to accepting the proprosed change, NON_COMPLIANT in the pre-enhancement stage, and the state of the same validation run as though the filled in values had been accepted, COMPLIANT in the post enhancement stage.

An assertion that the outcome was CURATED means that the data quality code has found an inconsistency between the data and an oracle and is proposing a correction based on the value from the oracle. If we have been careful in the interpretation of results returned from external authoritative services, and if those services themselves are accurate, then this proposed correction is probably reasonable. Proposed corrections of scientific names have posed the greatest challenge here. Scientific name authorities can vary in quality and, if selected, can return values that knowledgeable users consider to be incorrect.

An assertion that the outcome was UNABLE_DETERMINE_VALIDITY can result from a precondition for the evaluation not being met. Examples include scientific name validation for a record that lacks a dwc:scientificName and dwc:scientificNameAuthorship, or an orthogonal concern that the execution hasn't

yet untangled, or a failure of an external service to respond to a query from the data quality code.

An assertion that the outcome was UNABLE_CURATE corresponds to the sense of "SolveWithMoreData" in [Morris et al., 2013]. It is an assertion that the data quality control code found a problem but was not able to propose a solution to that problem. A composite record that contains textual locality information that describes one place on the surface of the Earth but a georeference for a different locality in a different country would produce this outcome. The georeference would be identified as inconsistent with the polygon for dwc:country, and if no transposition or sign change for the coordinates placed them near (e.g. within 20 km) of a georeference for the textual locality data as asserted by the geolocate service, FP-Akka would mark the outcome as UNABLE_CURATE. A human will need to evaluate this record and determine if the problem lies in the textual locality data, in the georeference, or in some other circumstance. An example of the latter is in the "Bernardo Assertion" in [Morris et al., 2013]), where identifications from three specimens were mis-transcribed by assigning locality data from three other specimens.

Presenting these outcome values to end users as short phrases posed a challenge. We initially presented them as icons accompanied by color coding of cells in a result spreadsheet. (See discussion at [FilteredPush, 2015].) Following repeated user feedback, we changed these to the text values, and then to brief human readable phrases shown in the "Outcome" column in Table 2. These labels capture the gist of the outcome (e.g. "CORRECT" carries "no change needed; looks good to us", where the "good to us" has the implication that the result may not be correct.)

The entries in Table 2 were derived[3] from actual data produced by FP-Akka applied to several specimen records at the Harvard Museum of Comparative Zoology(MCZ). Curation status outcome phrases such as these may be adequate for an analytical user of a large data set to include/exclude records that fit particular data quality criteria. They are, however, inadequate for a data curator assessing whether to apply the change proposed along with an outcome of CURATED to the database of record, or whether or not to expend effort to research further data needed to resolve a record with an outcome of UNABLE_CURATE. For these users, the details of the provenance in the list of sources consulted and the chain of assertions in the curation comment is critical.

In Table 2 we illustrate examples of QC assertions extracted from a spreadsheet produced from an actual dataset at the MCZ [4] processed by FP-Akka and subjected to the Java postprocessor that turns the JSON into a spreadsheet. The extractions are from several of the six sheets in the outputs of the postprocessor runs The rows of Table 2 are color coded[5] to signify one of the five outcome types as described in the next section. For ease of human consumption, each of the outcome types is rendered in natural language.

[3] See also Table 2 for examples derived from the postprocessor

[4] See Appendix A

[5] The precise color may vary slightly depending on the spreadsheet application in use and may find slight differences from the colors rendered in this paper. This is because the postprocessor is based on the Apache POI-HSSF platform (https://poi.apache.org/spreadsheet/index.html) for access to Microsoft Excel Format Files. POI-HSSF allows the spreadsheet reading program to choose the closest supported color to that requested by the postprocessor program.

Catalog Number	Validator	Outcome	Provenance
27366	EventDate	no change needed; looks good to us	Unable to construct eventDate from atomic fields. |eventDate is in ISO format |eventDate is consistent with modified date |Unable to get the Life span data of collector:Wayne P. Maddison |Unable to lookup a lifespan for the collector Wayne P. Maddison
27366	Scientific Name	we have proposed this change	Atomic fields for scientific name are blank, nothing to compare with scientific name. |Didn't find name in IPNI. |Found a name Sassacus vitis (Cockerell, 1894) which is in the same lexical group as the searched scientific name and claimed by GNI to be in IPNI but failed to find this name in IPNI. |No match found in IPNI with failover to GNI. ||The provided name: Sassacus vitis has a match in the GlobalNames Resolver |Didn't find name in IPNI. |Found a name Sassacus vitis (Cockerell, 1894) which is in the same lexical group as the searched scientific name and claimed by GNI to be in IPNI but failed to find this name in IPNI. |No match found in IPNI with failover to GNI. |Can't find the scientific name and authorship by searching the lexical group in GNI. |Got a valid result from GBIF checklistbank Backbone |The original SciName and Authorship are curated |Authorship: Author Dissimilar Similarity: 0.7333333333333333
62740	Scientific Name	no values present; we have proposed one	Atomic fields for scientific name are blank, nothing to compare with scientific name. |Didn't find name in IPNI. |Found a name Latrodectus Walckenaer, 1805 which is in the same lexical group as the searched scientific name and claimed by GNI to be in IPNI but failed to find this name in IPNI. |No match found in IPNI with failover to GNI. ||The provided name: Latrodectus has a match in the GlobalNames Resolver |Didn't find name in IPNI. |Found a name Latrodectus Walckenaer, 1805 which is in the same lexical group as the searched scientific name and claimed by GNI to be in IPNI but failed to find this name in IPNI. |No match found in IPNI with failover to GNI. |Can't find the scientific name and authorship by searching the lexical group in GNI. |Got a valid result from GBIF checklistbank Backbone |The original SciName and Authorship are curated |Authorship: Author Added Similarity: 0.0
62740	GeoRef	don't know	Both longitude and latitude are missing in the incoming Specimen-Record
27366	EventDate	there seems to be a problem, but we don't know how to solve it	dwc:eventDate does contain a value. |dwc:eventDate is not consistent with atomic parts (1985-06-06 ¡¿ [0][null][][][]) |dwc:verbatimEventDate parses to the same value as dwc:eventDate.

Table 2: Example QC Assertions. See also Appendix A

The Provenance column in Table 2 provides machine-produced provenance of the workflow. (It is denoted "Curation Comment" in the JSON rendering.) Notice that in most rows the provenance trace comprises several issues, separated by vertical bars. This arises from the presence in the JSON of a single Curation Comment element in the interface between the actor internals and the actor. The internal data quality control logic appends assertions to the Curation Comment as it works through the validation process for the provided elements of a record.

With FFDQ, comments in the provenance trace can be associated with individual assertion tests that make up the actor as a whole. A more structured representation of the provenance, as a graph of interdependent assertions that correspond to granular tests also makes it easier to identify the code implementing the tests.

The sequence of statements in the Provenance reflects the sequence of the internal logic of the quality control code (e.g. Figure 2). Not shown in the example data set is a common problem in integrated data sets. Namely, European dates are often shown as "day month year" whereas North American dates place month first. Best practice is to use ISO 8601 data representation [ISO, 2004] for dwc:eventDate. That said, if FP-Akka were examining multiple records at once, it might well glean enough information to speculate that European usage is in play. Indeed, deployments in FilteredPush nodes store the outcomes in MongoDB and could conceivably exploit these.

FP-Akka can consult one or more external and local services, depending on the workflow consulted and the configuration parameters provided. The list of services consulted in the response for each actor for each record indicates which services were invoked. Services include those below, although the release current at this writing, FP-Akka 1.6.0 [Kurator Wiki, 2015] does not consult them all.

- Tulane Geolocate http://www.museum.tulane.edu/geolocate/
- Scientific name services of Index Fungorum (IF) http://www.indexfungorum.org/
- The nomenclatural service from the International Plant Names Index (IPNI) http://www.ipni.org/
- Scientific name and taxonomy services from GBIF http://api.gbif.org/v1. (We consult the Backbone-Taxonomy, the code can invoke any names dataset in GBIF's checklist bank, but we currently only implement the Backbone Taxonomy)
- A service of our own, "Collecting Event Outlier Identification Service", which determines whether sequential collection events by the same collector on the same day are unreasonably far apart geographically, suggesting that there is an issue with the event time or location (Not currently enabled; nor actually a service, but is code for comparision within the a single dataset)
- A service of our own providing floral phenology data from the Flora of North America (FNA) [FNA, 2008] in order to compare the provided occurrence event date to the FNA flowering date range for records for which the flowering state if is recorded (and the occurrence Locality is in North America.) (Not currently enabled)
- Scientific name services of the Catalog of Life (COL)
- Scientific name services from the World Register of Marine Species (WoRMS) http://marinespecies.org/,
- Scientific name services from GlobalNames (ZooBank isn't turned on in current implementation)
- GNI as a failover service
- An agent first/last date service from the Harvard List of Botanists
- An agent first/last date service we have added to Symbiota.

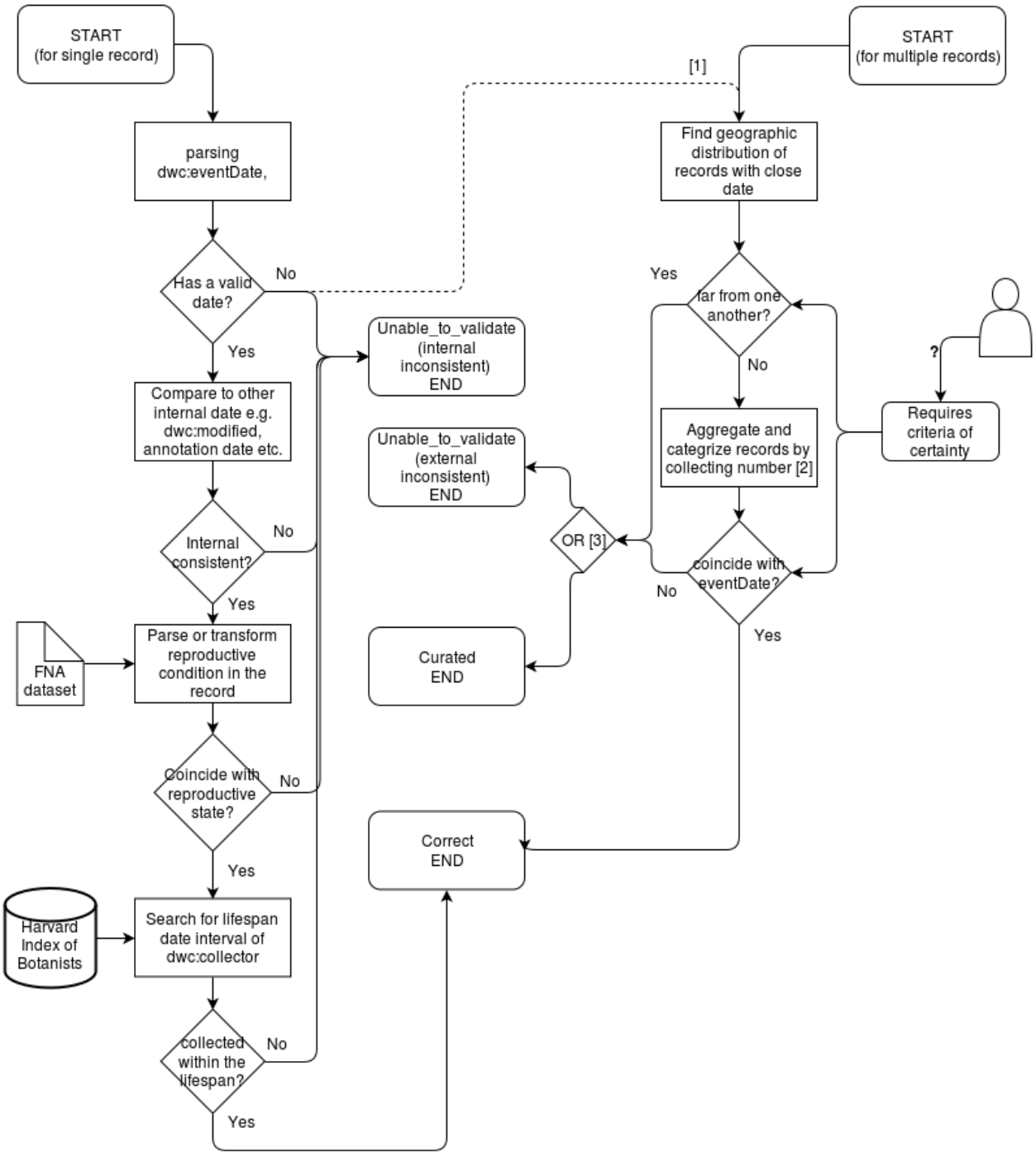

Figure 2: A flowchart of logic carried out by the DateValidator actor in the DwCa workflow.

In Kepler-Kuration, users were able to examine the curation status for an actor acting upon a record, examine a brief comment on that status, and build a detailed provenance graph of the flow of data through the workflow [Dou et al., 2012]. We observed that natural science collection managers and other biodiversity users of the workflow found the provenance graph too detailed about the Kepler provenance. They likewise found the very brief report added as columns to a single spreadsheet row for each record in a data set was not rich enough in domain assertions. We thus moved to a much more detailed per-record, per-actor data structure and began expanding the assertions made by the QC code in the course of the analysis. Presentation of this richer QC report presented the challenge described next, along with a solution modeled by a spreadsheet.

Reports expressed using FFDQ expand the single row per record format to include a list of all tests performed per record. The format of this spreadsheet models each record as a block of one or more rows of assertion test results per actor.

7 Why a spreadsheet?

FP-Akka is a Java-based workflow system to produce QC assertions, along with provenance for them. An intermediate structured data report is normally serialized as JSON. This format conveniently supports data embedded in JavaScript data for web applications, but not all interested parties have facilities, need, or desire for a web-based visualization of QC assertions. Furthermore, whereas JSON is rather programmer- and machine-friendly, it is hardly so for domain scientists. Appendix A discusses a complete small example of the JSON that is produced from a DwCa archive generated from a published taxonomic paper. FP-Akka can stand-alone, running on any machine that can run Java programs. A second stand-alone Java postprocessor program can convert the JSON output to a spreadsheet conforming to the Microsoft Excel xls format. The result is the focus of this paper. The spreadsheet contains cells showing the proposed QC suggestions or indications that FP-Akka cannot offer anything. The corresponding cells are color coded to signify the nature of the outcome. For each record, data attributes that have been examined and found fit (as determined by the workflow) are in green cells, but even for these, the provenance for this finding is reported. This spreadsheet allows data managers or users to evaluate fitness for use and take action with their own tools, no matter whether or not the FP-Akka JSON output can be consumed by their data management or scientific application tools.

8 Continuous Quality Control, Speed and Parallelization

Taxon or occurrence data is never final. Providing for Quality Control is a continuous ongoing enterprise. Natural science collections are living and breathing libraries of our knowledge of biological diversity. Scientific names change based on emerging science and according to nomenclatural rules. But other kinds of ongoing changes arise due to curatorial practices (e.g. community adoption of data serialization standards) funding sources (e.g. for digitization of dark data or for community georeferencing of localities). Yet others arise from modern software architecture practices required for big data sets, particularly those increasing rapidly in size, such as social media data (e.g.[Cai and Zhu, 2015, Immonen et al., 2015]) and sensor data (e.g. [Campbell et al., 2013])

An ongoing issue faces users of FP-Akka in the form of some non-deterministic record ordering. Mainly due to parallelization and variation in response time of remote services, production of output records can complete in an order not perfectly correlated to their input order. Consequently, two subsquent executions of FP-Akka may not have the same output order. Moreover, a remote service might timeout differently in the two, and the results may lead to different assertions. Indeed, even without timeout, the remote service

may have simply offered changed data, possibly also resulting in different FP-Akka output assertions. This is not as grim as it seems, because many large datasets suffer from systematic errors, e.g. collector name misspellings, latitude and longitude reversed, date ambiguities such as mentioned earlier, etc. Thus, an input set of several hundred thousand occurrences may result in 10,000 QC assertions, many of which can be fixed with only dozens of curations. Such situations may be common in early stages of specimen digitization projects.

A critical next step for the Kurator project is to render data quality reports to show unique instances of errors, where unique instances often correspond with values in a single value in a row in a relational database table – a single value that has been expanded out to multiple repeating values in the exported flat DarwinCore view of the data. Using FFDQ, a report like this could be modeled as a multi-record, i.e. dataset, level assertion. Another example using FFDQ is a report that shows multi-record level measures reporting on the percentage of records that satisfy some set of single record validations before and after an amendment stage in a workflow. In FP-Akka, our reports were compilations of record-level (single record in the FFDQ sense) assertions, and dataset level assertions are treated as statistical compilations over those record-level assertions. FFDQ is more general and richer, in that it can describe dataset level measures that aggregate over tests on single records.

9 DwCa Workflow

The default workflow supported by FP-Akka is named *DwCa* following the acronym for Darwin Core Archive, a typical source of occurrence data for this workflow. Figure 3 shows DwCa workflow in FP-Akka produced from YesWorkflow markup of the workflow class [McPhillips et al., 2015]. Green boxes represent actors in the workflow, yellow boxes represent data records transferred (as messages in the context of Akka) between the actors. Inputs to the system are circles and arrows on the outer border. Read this diagram starting with the load of the inputFile by the CSVReader actor. This workflow incorporates a simple rate-throttling mechanism for PullRequestor actors positioned after all the potentially slow actors that invoke external services. Each time such an actor receives a geoRefValidatedRecord, the actor sends a message to the CSVReader, which in turn requests that another record be sent down the pipeline.

Performance testing suggests that refactoring our existing actors to be more granular can give us even more of a performance boost, particularly in the case of actors that make calls to external services. The increased granularity allows FP-Akka to assign multiple processors when they are available and also consult a different actor should one block if one of *its* remote services blocks. In addition, increased granularity allows more flexibility in code reuse.

10 CSV workflow

FP-Akka also supports a second use - evaluating and cleaning taxonomic authority tables in natural science collection databases in comparison with external authorities. FP-Akka-workflowstarter can be invoked using a "CSV" workflow with a switch that runs a workflow that includes only a CSV reader, the scientific name validation actor, and a CSV writer. This workflow is designed to take a dump of data from a taxon authority table, run it against a specified authority, and produce a simple flat report from which records with particular outcomes can be selected to either send to a taxonomically knowlegable collections management staff member to evaluate and correct in the database, or which can be converted (easily with grep/sed/awk) into sql update commands that can be fired against the taxonomic authority table in the database of record. Figure 4 contrasts two ways in which FP-Akka can be incorporated into data quality processes by curators of a natural science collections database of record.

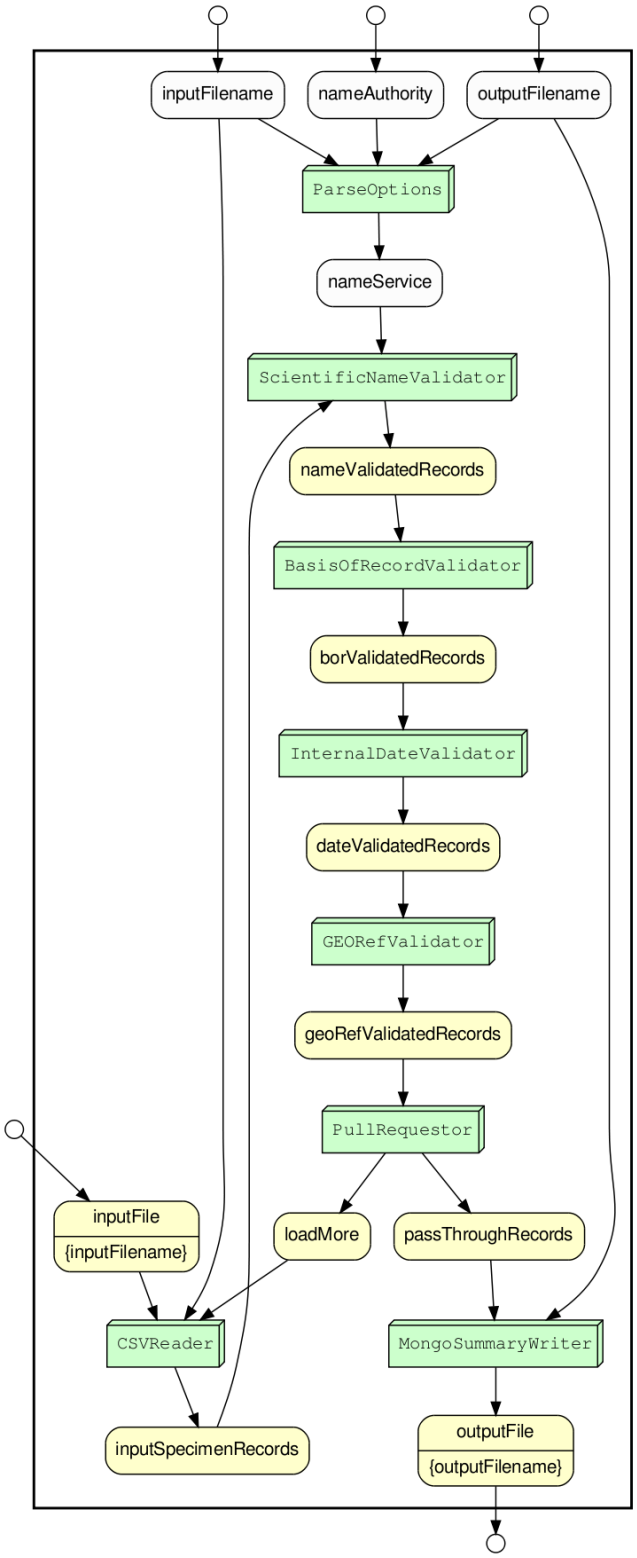

Figure 3: DwCa Workflow

An alternative implementation that serializes the concepts in an RDF graph and uses SPARQL queries to generate reports as some set of assertions enables us to write a single postprocessor capable of generating a wide variety of reports. In some cases, depending on a particular use case, users may only be interested in a subset of the tests performed by an actor. In other cases, a user may only be interested in a subset of the outcomes such as the actionable items only.

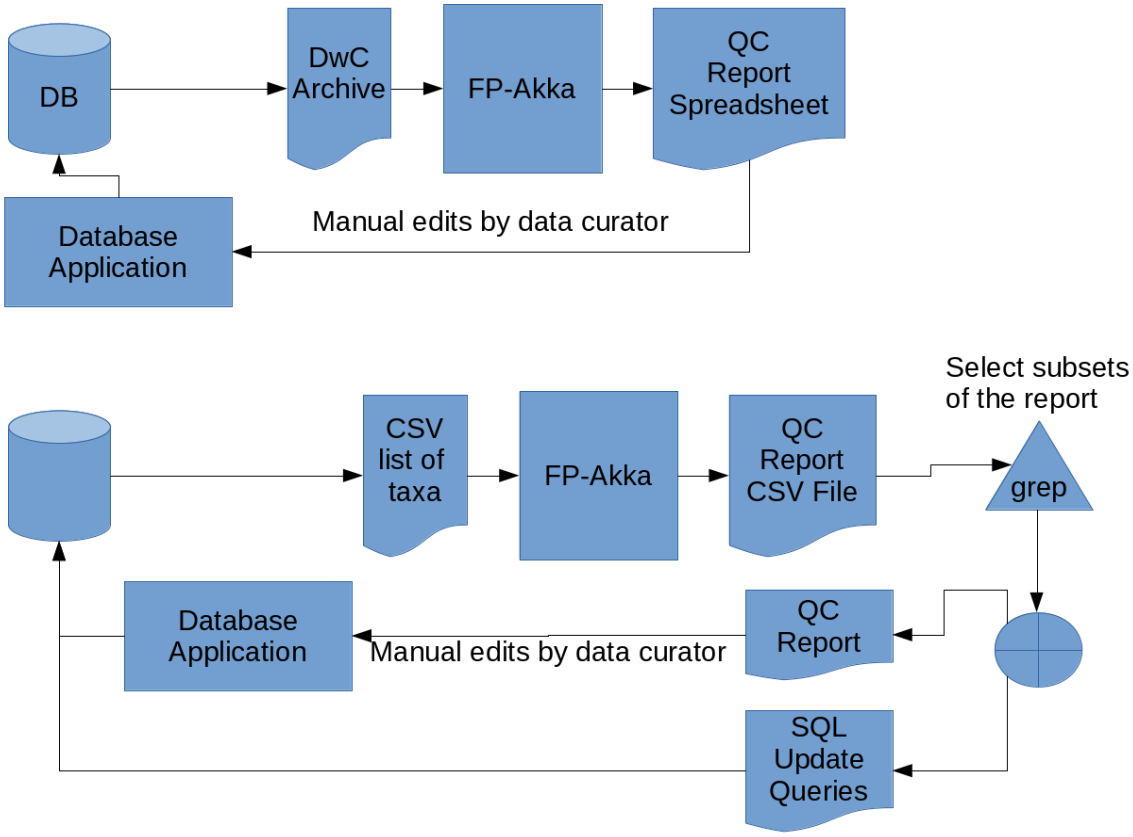

Figure 4: Incorporation of FP-Akka into data quality procedures in a natural science collection. Thw upper diagram depicts typical use of the DwCa workflow; the lower diagram deoicts alternative use cases supported by the CSV workflow. Both are described in [Kurator Wiki]

11 QC of the QC software

A QC assertion, especially a systematically occurring one, that is inexplicable to an expert human user may rightly raise the question of whether the input data has exposed a QC issue in the QC software itself. FP-Akka itself is defended in part by standard software engineering QC mechanisms, such as unit testing [Fraser et al., 2003] with the JUnit framework [JUnit, 2015], logging application runtime behavior (especially upon timeout of remote services), etc. Some, but not all, of these mechanisms reflect their test results to the console at the time FP-Akka is executed. They are largely targeted at programmers, and their results are not generally available in the post-processor spreadsheet. That kind of QC is outside the

scope of this paper. We have, however, over several years, repeatedly sent the output of FP-akka runs on datasets (especially those of the Symbiota Collections of Arthropods Network [SCAN TCN website])

Some defects in the software that we have identified through feedback from users of quality control reports include systematic errors. One example iss that all marine localities were flagged as errors due to a developer's interpretation of the phrase 'on the Earth's surface' to mean on land rather than latitude between +/- 90 degrees. Some subtle problems evaded unit tests, one of which is described below. Most of the issues that we are seeing now in feedback from users are yet another class of problem. These are not defects in the software itself, but unexpected values present in the responses from nomenclatural authorities. For example, in a recent batch of work that involved disambiguating which member of the Sowerby family should be attributed to which scientific names in the taxonomic authority file for Malacology in the Harvard Museum of Comparative Zoology, the collection management staff member to whom an FP-Akka report was given noted that the authorship for the genus Stilifer of 'Broderip [in Broderip & Sowerby I], 1832' seemed odd. FP-Akka, consulting WoRMS, had taken the input dwc:scientificName=Stilifer, dwc:scientificNameAuthorship=Broderip and Sowerby, 1832, had returned the record as curated, with the curation comment "Found plausible match in WoRMS: Specifying Which Sowerby, Year Exact" and a guid for the match in WoRMS of urn:lsid:marinespecies.org:taxname:205197. The user was expecting, for their purpose, for the formulation of the correction to be Stilifer Broderip and Sowerby, 1832 curated as Stilifer Broderip, 1832, instead of the formulation presented by WoRMS. We have seen similar issues in other zoological authorities in which the facts presented from the authority are correct, but the format presented from the authority does not match the expectations of the user of the FP-Akka report.[6]

We next describe serious misbehavior of FP-Akka in response to curious, surely unintended, georeference in the data described in the DwC Archive served by [Čandek et al., 2015] as of November 12, 2015. The occurrence data in that DwCa had all decimalLatitude and decimalLongitude 10 times the intended values, e.g. latitude 458.797 (rather than 45.87556) and longitude 139.468 (rather than 13.948889) for the Slovenian town of Budanje. One of the georeference errors that FP-Akka checks is that error in decimal latitude and longitude values. The expected response for these input values would be either curation to the divided-by-10 values, or a marker that the record was in error. In many cases, however, FP-Akka asserted that these numerically invalid input values are correct, with provenance suggesting that the widely used Tulane Geolocate service was accepting such values and—inspection showed—was querying and getting answers for values as though divided by 10. For the example, the FP-Akka output signaled "no change needed" for what is clearly a ridiculous result. However, in this case, the spreadsheet provenance column carried the text

"found 1 possible georeferences with Geolocate engine:GLC:4.95|U:1.01374|eng:1.0 |BUDANJE score:82 45.875556 13.948889 km:0 |Original coordinates are near (within georeference error radius or 20.0 km) the georeference for the locality text from the Geolocate service. Accepting the original coordinates."

This suggested erroneously that the original coordinates are acceptable, while at the same time indicating that the Geolocate service, invoked with locality Budanje, returned reasonable coordinates.

Investigation of our georeference validation code located a subtle defect in a method that finds the distance between two points on the earth independent of latitude, and a unit test that was correct, but of inadequate scope to catch this defect. Examination of the code revealed two places where the invocation of the trigonometric functions *sin* and *cos* neglected to convert latitude and longitude to radians as assumed by the Java Math library. Consider comparing the points $(458.797, 139.468)$ and $(45.87556, 13.9468)$ For the first of these points, the spherical trigonometry function *haversine* has a factor containing the erroneous $cos(458.797)$ hence was driven negative and so was the term containing that factor. To get a great circle distance in kilometers between the two points, this term is fed to the square root function *sqrt* thence to the arctangent function $atan2$ with corresponding erroneous longitude. But the square root of a

[6] See http://marinespecies.org/aphia.php?p=taxdetails&id=205197

negative number is not a real number, so *sqrt* returns the Java numeric constant NaN ("Not a Number"). Subsquently NaN is passed to *atan*2 which returns 0 on these NaN's to compute the great circle distance. Thus $(458.797, 139.468)$ and $(45.87556, 13.9468)$ are the same point in the face of our coding error. The unit test asked the method to calculate the distance from the equator to 1 degree North, and the distance from one point on the equator to one degree West of that point. In both cases, a zero value was passed in to be multiplied by the results of the sine and cosine functions that were handed degrees instead of radians, causing the error to go away and for these narrow cases, for a correct result to be returned. A unit test (now present) that asked for the distance between two arbitrary points on the surface of the earth, neither on the equator or Greenwich meridian, would have failed. Thus, the issue had nothing to do with the Geolocate service. The current release of FP-Akka has correct code.

A key point in this realization of where defects may lie is expectation management for the end users - things we report as errors may be defects in their data, defects in our code, or defects in the authorities our workflows are consulting. Users to whom the data quality report is presented without this understanding may observe what they percieve as errors in the data quality report. They might dismiss the reporting framework as defective, unless they understand that their data may expose errors in the framework that the developers can readily fix, or that the errors are in the external authoritative sources that the software consulted.

FFDQ provides more descriptive metadata about the tests performed through the concepts Specification, Criteria, Enhancement and Dimension. This information can be included or linked to results in the spreadsheet to provide additional background information about the expected behavior of test. Also related is work done to standardize tests since the assertions can be linked to a standard definition of tests that users will be familiar with. In FP-Akka, these metadata are buried in (marginally) human readable form in curation comments, rather than being exposed as metadata that can easily be tied to elements in an report. FFDQ allows for a report to be provided about the tests themselves. A typical user question is "what is this test doing" or "what fields are required for this test". FP-Akka's large and complex actors bury this information in comments. FFDQ allows provision of this information as queriable test metadata.

12 Acknowledgements

We thank SCAN curators and participants in Kurator workshops who have reviewed FP-Akka QC reports on their data, Brendan Haley and Adam Baldinger in the MCZ who have reviewed FP-Akka QC reports on their data, and Susan Morris for comments on the manuscript.

A Appendix A MCZData

Attached is a spreadsheet
iot = https://github.com/kurator-org/FP-Akka-Manuscript/blob/master/data_in_and_data_out_v2.xls

The spreadsheet *iot* has three tables, which together illuminate how the data in Table 2 is derived from two underlying records in the MCZ specimen database bearing catalog numbers 27366 and 6270740, respectively. That said, Table 2 may be taken on faith without impediment to reading the paper. Otherwise, it is best to have some familiarity with
http://wiki.datakurator.org/wiki/FP-Akka_User_Documentation, especially
http://wiki.datakurator.org/wiki/FP-Akka_User_Documentation#Workflow_2.

The first table in *iot* is a CSV dump of the aforementioned records, with artificial repetition for convenience in consideration of discussion in the next two tables in *iot*.

The second table in *iot* corresponds to the output of the post processor as described in
http://wiki.datakurator.org/wiki/FP-Akka_User_Documentation#Post_processing

The third table in *iot* is crafted from parts of the second *iot* table, to provide the data for Table 2 in the main body of the paper.

The following odd part of the Provenance in the bottom row of Table 2 is produced by code in FP-Akka and indicates a failure to curate the data. In this case FP-Akka was unable to fill in a value for event date from the atomic fields (startDayOfYear, endDayOfYear, year, month, day) because these values are empty or null in the input data:

"dwc:eventDate is not consistent with atomic parts (1985-06-06 ¡¿ [0][null][][][])"

References

ABCD. Task Group on Access to Biolgical Collection Data, 2015. URL http://www.bgbm.fu-berlin.de/TDWG/CODATA/default.htm.

Jorge A. Ahumada, Carlos E. F. Silva, Krisna Gajapersad, Chris Hallam, Johanna Hurtado, Emanuel Martin, Alex McWilliam, Badru Mugerwa, Tim O'Brien, Francesco Rovero, Douglas Sheil, Wilson R. Spironello, Nurul Winarni, and Sandy J. Andelman. Community structure and diversity of tropical forest mammals: data from a global camera trap network. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1578):2703–2711, 2011. ISSN 0962-8436. doi: 10.1098/rstb.2011.0115.

Assn. of Systematics Collections. An information model for biological collections, October 1992. URL http://cool.conservation-us.org/lex/datamodl.html.

BDQ. Biodiversity data quality (bdq) interest group, 2017. URL https://github.com/tdwg/bdq. [Online; Accessed 2017-11-23].

BioCASE. Biological Collection Access Services, 2015. URL http://www.biocase.org/.

M de L. Brooke. Why museums matter. *Trends in Ecology & Evolution*, 15(4):136–137, 2000. ISSN 0169-5347. doi: 10.1016/S0169-5347(99)01802-9. URL http://dx.doi.org/10.1016/S0169-5347(99)01802-9.

Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, 2015. doi: http://doi.org/10.5334/dsj-2015-002.

John L. Campbell, Lindsey E. Rustad, John H. Porter, Jeffrey R. Taylor, Ethan W. Dereszynski, James B. Shanley, Corinna Gries, Donald L. Henshaw, Mary E. Martin, Wade M. Sheldon, and Emery R. Boose. Quantity is nothing without quality: Automated qa/qc for streaming environmental sensor data. *BioScience*, 63(7):574–585, 2013. doi: 10.1525/bio.2013.63.7.10. URL http://bioscience.oxfordjournals.org/content/63/7/574.abstract.

Arthur D Chapman. Principles of Data Quality. Technical report, Global Biodiversity Information Facility, Copenhagen, 2005. URL http://circa.gbif.net/irc/DownLoad/kjeYAKJSmRGFqwAaUY4x8KZ1jH4pYxtv/F37w1fUI4ROAgTiySEZttf0yRVSbNGn/DataQuality.pdf.

CLO. Understanding the ebird review and data quality process, 2012. URL http://help.ebird.org/customer/en/portal/articles/1055676-understanding-the-ebird-review-and-data-quality-process.

R.A. Creighton and J.J. Crockett. Selgem: A system for collection management. *Information Systems Innovations*, 2(3):1–26, 1971.

Lei Dou, Daniel Zinn, Timothy McPhillips, Sven Köhler, Sean Riddle, Sean Bowers, and Bertram Ludäscher. Scientific workflow design 2.0: Demonstrating streaming data collections in Kepler. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, pages 1296–1299. IEEE, 2011.

Lei Dou, G Cao, Paul J Morris, Robert A Morris, Bertram Ludäscher, James A Macklin, and James Hanken. Kurator: A Kepler Package for Data Curation Workflows. *Procedia Computer Science*, 9(0):1614–1619, 2012. doi: 10.1016/j.procs.2012.04.177. URL http://www.sciencedirect.com/science/article/pii/S1877050912002980.

ECMA Iternational. Final draft of the TC39 "The JSON Data Interchange Format" standard - ECMA-404.pdf, 2013. URL http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf.

FilteredPush. Embedding Kepler — FilteredPush 6.1.3 Result types and symbols, 2015. URL http://wiki.filteredpush.org/w/index.php?title=Embedding_Kepler&oldid=7760#Result_types_and_symbols. [Online; accessed 20-November-2015].

FilteredPushWiki. FilteredPushWiki, 2016. URL http://wiki.filteredpush.org/wiki/FilteredPush. [Online; Accessed 2017-11-23].

FishNet web site. About FishNet, 2013. URL http://www.fishnet2.net/aboutFishNet.html. [Online; Accessed 2016-11-26.

FNA. Flora of North America, 2008. URL http://floranorthamerica.org/.

FP web site. FilteredPush Web, 2016. URL http://wiki.filteredpush.org/wiki/. [Online; Accessed 2016-03-09].

Steven Fraser, Dave Astels, Kent Beck, Barry Boehm, John McGregor, James Newkirk, and Charlie Poole. Discipline and practices of tdd: (test driven development). In *Companion of the 18th Annual ACM SIGPLAN Conference on Object-oriented Programming, Systems, Languages, and Applications*, OOPSLA '03, pages 268–270, New York, NY, USA, 2003. ACM. ISBN 1-58113-751-6. doi: 10.1145/949344.949407. URL http://doi.acm.org/10.1145/949344.949407.

Jennifer K. Frey. Inferring species distributions in the absence of occurrence records: An example considering wolverine (*Gulo gulo*) and canada lynx (*Lynx canadensis*) in New Mexico. *Biological Conservation*, 130(1):16 – 24, 2006. ISSN 0006-3207. doi: http://dx.doi.org/10.1016/j.biocon.2005.11.029. URL http://www.sciencedirect.com/science/article/pii/S0006320705005264.

GBIF. Darwin Core Archive Assistant, 2013. URL http://tools.gbif.org/dwca-assistant. [Online; accessed 27-November-2016].

GBIF. Global biodiversity information facility, 2015. URL http://www.gbif.org.

Catherine H. Graham, Simon Ferrier, Falk Huettman, Craig Moritz, and A. Townsend Peterson. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, 19(9):497 – 503, 2004. doi: {http://dx.doi.org/10.1016/j.tree.2004.07.006}. URL http://www.sciencedirect.com/science/article/pii/S0169534704002034.

Harvard University Herbaria. Index of Botanists, 2013. URL http://kiki.huh.harvard.edu/databases/botanist_index.html.

Elspeth Haston, Robert Cubey, Martin Pullan, Hannah Atkins, and David Harris. Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. *ZooKeys*, 209:93–102, jul 2012. doi: 10.3897/zookeys.209.3121. URL http://dx.doi.org/10.3897/zookeys.209.3121.

HerpNET final report. HerpNET Final Report Highlights, 2009. URL http://herpnet.org/Gazetteer/finalreport.html. [Online; Accessed 2016-03-08].

HISCOM. Hispid 5 for hispid users, December 2013. URL http://hiscom.rbg.vic.gov.au/wiki/HISPID_5_for_HISPID_Users.

iDigBio. NSF ADBC Program Information | iDigBio, 2011. URL https://www.idigbio.org/content/nsf-adbc-program-information.

iDigBio. iDigBio Home, 2015. URL https://www.idigbio.org/home.

IETF. RFC 4180 - Common Format and MIME Type for Comma-Separated Values (CSV) Files, 2005. URL https://tools.ietf.org/html/rfc4180.

IF. Index Fungorum, 2015. URL http://www.indexfungorum.org/.

A. Immonen, P. Paakkonen, and E. Ovaska. Evaluating the quality of social media data in big data architecture. *Access, IEEE*, 3:2028–2043, 2015. ISSN 2169-3536. doi: 10.1109/ACCESS.2015.2490723.

IPNI. The International Plant Names Index, 2012. URL http://ipni.org/.

ISO. ISO 8601:2004, December 2004. URL http://www.iso.org/iso/home/standards/iso8601.htm.

JUnit, 2015. URL http://junit.org/.

Kurator-FFDQ. URL https://doi.org/10.5281/zenodo.192186. FFDQ Framework DOI: 10.5281/zenodo.192186.

Kurator-FFDQAPI. URL https://doi.org/10.5281/zenodo.891415. FFDQ API DOI: 10.5281/zenodo.891415.

Kurator Project. Fp-akka post processing, 2015. URL http://wiki.datakurator.net/web/FP-Akka_User_Documentation#Post_processing.

Kurator Wiki. FPAkka User Documentation, 2015. URL http://wiki.datakurator.net/w/index.php?title=FP-Akka_User_Documentation&oldid=451. [Online; accessed 14-Dec-2016; FP-Akka Release 1.6.0].

Kurator Wiki. Kurator Wiki, 2016. URL http://wiki.datakurator.org/wiki/Kurator. [Online; Accessed 2016-12-16].

Frank E. Lutz. *The Variation and Correlations of Certain Taxonomic Characters of Gryllus*. Number No. 101 in Carnegie Institution of Washington Publication. Carnegie Institution of Washington, Washington, D.C., 1908. Carnegie Institution of Washington, Publication No. 101.

Timothy M. McPhillips, Paul J. Morris, David A. Lowery, Qian Zhang, John Wieczorek, and Allan Koch Veiga. Kurator-akka. URL https://github.com/kurator-org/kurator-akka.

Timothy M. McPhillips, Tianhong Song, Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, Kyle Bocinsky, Yang Cao, Fernando Chirigati, Saumen C. Dey, Juliana Freire, Deborah N. Huntzinger, Christopher Jones, David Koop, Paolo Missier, Mark Schildhauer, Christopher R. Schwalm, Yaxing Wei, James Cheney, Mark Bieda, and Bertram Ludäscher. YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. *International Journal of Digital Curation*, 10(1):298–313, 2015. doi: 10.2218/ijdc.v10i1.370. URL http://ijdc.net/index.php/ijdc/article/view/370/0.

Inc MongoDB. The MongoDB 3.0 Manual — MongoDB Manual 3.0, March 2015. URL https://docs.mongodb.org/manual/.

Paul J. Morris. Relational Database Design and Implementation for Biodiversity Informatics. *PhyloInformatics*, 7:1–66, 2005. URL http://www.athro.com/general/Phyloinformatics_7_85x11.pdf#.

Robert A Morris, Lei Dou, James Hanken, Maureen Kelly, David B. Lowery, Bertram Ludäscher, James A. Macklin, Paul J. Morris, Robert A. Morris Mail, Lei Dou, James Hanken, Maureen Kelly, David B. Lowery, Bertram Ludäscher, James A. Macklin, and Paul J. Morris. Semantic Annotation of Mutable Data. *PLoS ONE*, 8(11):1–29, 2013. doi: doi:10.1371/journal.pone.0076093. URL http://dx.plos.org/10.1371/journal.pone.0076093.

OBIS. Ocean Biogeographic Information System, 2011. URL http://www.iobis.org/.

ORNIS web site. ORNIS, 2016. URL http://www.ornisnet.org/. [Online; Accessed 2016-03-08].

G. S. (George Stanley) Radford. *The control of quality in manufacturing*. The Ronald Press Company, New York, 1922. URL http://archive.org/details/controlofquality00radf.

Tim Robertson, Markus Döring, John Wieczorek, Renato De Giovanni, and David Vieglais. Darwin Core Text Guide, June 2015. URL http://rs.tdwg.org/dwc/terms/guides/text/.

SCAN TCN website. Home — Symbiota Collections of Arthropods Network, 2016. URL http://scan1.acis.ufl.edu/. [Online; Accessed 2016-11-07].

Stanwyn G. Shetler. Demythologizing biological data banking. *Taxon*, 23(1):71–100, 1974. ISSN 00400262. URL http://www.jstor.org/stable/1218091.

Walter A Shewhart. *Statistical Methods from the Viewpoint of Quality Control*. The Graduate School, U.S. Department of Agriculture, 1939. Republished by Dover Publications.

Vincent Smith and Vladimir Blagoderov. Bringing collections out of the dark. *ZooKeys*, pages 1–6, jul 2012. doi: 10.3897/zookeys.209.3699. URL http://dx.doi.org/10.3897/zookeys.209.3699.

Smith et al. No specimen left behind: mass digitization of natural history collections, 2012. URL http://zookeys.pensoft.net/browse_journal_issue_documents?issue_id=361.

Tianhong Song, Sven Köhler, Bertram Ludäscher, James Hanken, Maureen Kelly, David Lowery, James A. Macklin, Paul J. Morris, and Robert A. Morris. Towards Automated Design, Analysis and Optimization of Declarative Curation Workflows. *International Journal of Digital Curation*, 9(2), 2014. doi: 10.2218/ijdc.v9i2.337.

Barbara Stein and John Wieczorek. Mammals of the world: Manis as an example of data integration in a distributed network environment. *Biodiversity Informatics*, 1(0), 2004. ISSN 15469735. doi: 10.17161/bi.v1i0.7. URL https://journals.ku.edu/index.php/jbi/article/view/7.

B L Sullivan, C L Wood, M J Iliff, R E Bonney, D Fink, and S Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 2009.

Mary E. Sunderland. Computerizing natural history collections. *Endeavour*, 37(3):150 – 161, 2013. ISSN 0160-9327. doi: http://dx.doi.org/10.1016/j.endeavour.2013.04.001. URL http://www.sciencedirect.com/science/article/pii/S0160932713000227.

TDWG. TDWG: History, May 2007. URL http://www.tdwg.org/about-tdwg/history/.

TDWG. Darwin Core Terms: A quick reference guide, June 2015. URL http://rs.tdwg.org/dwc/terms/index.htm.

Typesafe Inc. Akka, 2015. URL http://akka.io/.

A K Veiga, A M Saraiva, A D Chapman, P J Morris, C Gendreau, D Schigel, and T J Robertson. A conceptual framework for quality assessment and management of biodiversity data. *PLOS ONE*, 12(6):1–20, 06 2017. doi: 10.1371/journal.pone.0178731. URL https://doi.org/10.1371/journal.pone.0178731.

VertNet. VertNet-About, 2015. URL http://www.vertnet.org/about/about.html.

Richard Y. Wang. A product perspective on total data quality management. *Commun. ACM*, 41(2):58–65, February 1998. ISSN 0001-0782. doi: 10.1145/269012.269022. URL http://doi.acm.org/10.1145/269012.269022.

John Wieczorek. MaNIS Home, 2015. URL http://manisnet.org/. [Online; Accessed 2016-03-08].

John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, 7(1):e29715, January 2012. doi: 10.1371/journal.pone.0029715. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0029715`.

J.H. Wiersema, J. McNeill, N. Turland, F.R. Barrie, W.R. Buck, V. Demoulin, W. Greuter, D.L. Hawksworth, P.S. Herendeen, S. Knapp, K. Marhold, J. Prado, W.F. Prud'homme van Reine, and G.F. Smith, editors. *International Code of Nomenclature for algae, fungi and plants (Melbourne Code) adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011*. Koeltz Scientific Books, 2012.

Wikipedia. Grep — wikipedia, the free encyclopedia, 2015. URL `https://en.wikipedia.org/w/index.php?title=Grep&oldid=684547908`. [Online; accessed 19-October-2015].

Edward O. Wiley and A. Townsend Peterson. Biodiversity and the Internet: Building and Using the Virtual World Museum. In Arno Scharl, editor, *Environmental Online Communication*, Advanced Information and Knowledge Processing, pages 91–99. Springer London, 2004. ISBN 978-1-84996-913-0 978-1-4471-3798-6. URL `http://link.springer.com/chapter/10.1007/978-1-4471-3798-6_11`. DOI: 10.1007/978-1-4471-3798-6_11.

Klemen Čandek, Matjaž Gregorič, Rok Kostanjšek, Holger Frick, Christian Kropf, and Matjaž Kuntner. Corrigendum: Targeting a portion of central european spider diversity for permanent preservation. *Biodiversity Data Journal*, 3:e4301, jan 2015. doi: 10.3897/BDJ.3.e4301. URL `http://dx.doi.org/10.3897/BDJ.3.e4301`.