EMuView: Field Museum Database and Metadata Visualization

Kartik Adur, Richard Higgins, Jingsha Luo, Joshua Quick, Wensi Wang

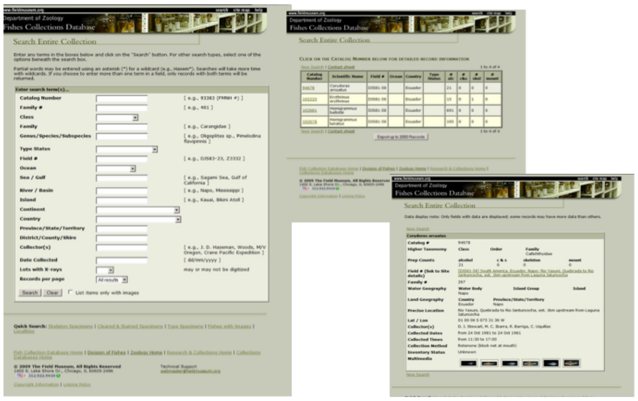
Image: http://technology.fieldmuseum.org/our-work

Clients

The Field Museum IT staff: Sharon Grant, Technology Liaison to Science; Katherine Webbink, Information Systems Specialist; Marc Lambruschi, Data Migration Technician

Introduction

Since 2002 The Field Museum of Natural History has been converting its collection databases to KE Software's EMu (Museum Collection Management System) across all four collection-based departments: Botany, Zoology, Anthropology, and Geology. More recently, they have been working on moving the old Web interfaces for the EMu system to a Drupal-based system. The Drupal transition enables the Museum to make available online a much larger fraction of the Museum's 24 million specimens, including entire collections databases (such as the Fishes collection, below) publicly searchable. [1]



Department of Zoology
Fishes Collections Database

Search Entire Collection

Enter any terms in the boxes below and click on the "Search" button. For other search types, select one of the options beneath the search box.

Partial words may be entered using an asterisk (*) for a wildcard (e.g., Hasem*). Searches will take more time with wildcards. If you choose to enter more than one term in a field, only records with both terms will be returned.

Enter search term(s)...

Catalog Number	[e.g., 93383 (FMNH #)]
Family #	[e.g., 481]
Class	
Family	[e.g., Carangidae]
Genus/Species/Subspecies	[e.g., Oligoplites sp., Pimelodina flavipinnis]
Type Status	
Field #	[e.g., DJS83-23, Z3332]
Ocean	
Sea / Gulf	[e.g., Sagami Sea, Gulf of California]
River / Basin	[e.g., Napo, Mississippi]
Island	[e.g., Kauai, Bikini Atoll]
Continent	
Country	
Province/State/Territory	
District/County/Shire	
Collector(s)	[e.g., J. D. Haseman, Woods, M/V Oregon, Crane Pacific Expedition]
Date Collected	[dd/mm/yyyy]
Lots with X-rays	may or may not be digitized
Records per page	All results

Search | Clear | □ List items only with images

Quick Search: Skeleton Specimens | Cleared & Stained Specimens | Type Specimens | Fishes with Images | Localities

Fish Collection Database Home | Division of Fishes | Zoology Home | Research & Collections Home | Collections Databases Home

© 2009 The Field Museum, All Rights Reserved
1400 S. Lake Shore Dr., Chicago, IL 60605-2496
Technical Support webmaster@fieldmuseum.org

Copyright Information | Linking Policy

Department of Zoology
Fishes Collections Database

Search Entire Collection

CLICK ON THE CATALOG NUMBER BELOW FOR DETAILED RECORD INFORMATION

New Search | Contact sheet — 1 to 4 of 4

Catalog Number	Scientific Name	Field #	Ocean	Country	Type Status	# alc	# c&s	# skel	# mount
94678	Corydoras arcuatus	DJS81-38		Ecuador		21	0	0	0
102223	Erythrinus erythrinus	DJS81-38		Ecuador		13	0	1	0
102681	Hemigrammus bellottii	DJS81-38		Ecuador		691	0	0	0
102678	Hemigrammus lunatus	DJS81-38		Ecuador		105	0	0	0

New Search | Contact sheet — 1 to 4 of 4

Export up to 2000 Records

Department of Zoology
Fishes Collections Database

Search Entire Collection

Data display note: Only fields with data are displayed; some records may have more data than others.

New Search

Corydoras arcuatus

Catalog #	94678
Higher Taxonomy	Class Order Family Callichthyidae
Prep Counts	alcohol 21, c & s 0, skeleton 0, mount 0
Field # (link to Site details)	DJS81-38 South America, Ecuador, Napo: Rio Yasuni, Quebrada to Rio Jatuncocha, est. 1km upstream from Laguna Jatuncocha
Family #	267
Water Geography	Water Body Napo, Island Group, Island
Land Geography	Country Ecuador, Province/State/Territory Napo
Precise Location	Rio Yasuni, Quebrada to Rio Jatuncocha, est. 1km upstream from Laguna Jatuncocha
Lat / Lon	01 00 06 S 075 31 36 W
Collector(s)	D. J. Stewart, M. C. Ibarra, R. Barriga, C. Ugullas
Collected Dates	from 24 Oct 1981 to 24 Oct 1981
Collected Times	from 11:30 to 17:00
Collection Method	Rotenone (block net at mouth)
Inventory Status	Unknown
Multimedia	

New Search

Figure 1: Drupal database interface (Grant 2011, slide 20).

Visualization Demo ella.ils.indiana.edu/~kadur/IVMOOCFinal/
Project Documentation github.com/rshiggin/IVMOOC2015-FMNH

This project was completed as part of IVMOOC 2015, a data visualization course based at Indiana University, Bloomington.

Authors contact information:

- *Kartik Adur, kadur@indiana.edu*
- *Richard Higgins, rshiggin@indiana.edu*
- *Jingsha Luo, luojin@umail.iu.edu*
- *Joshua Quick, jdquick@indiana.edu*
- *Wensi Wang, wenswang@imail.iu.edu*

Because of the scale of this work, the technology staff of The Field Museum has pursued ways of visualizing forms of activity and interaction in the EMu system. They asked us to use metadata from the EMu system — essentially audit or log records collected for each database transaction — to highlight and visualize the scope and range of activity in their collection management system. Such demonstrations can be used to show museum staff and other stakeholders the breadth of database activity, reveal what objects, categories, and users are most active over time and in specific time intervals, and identify important patterns among different departments, modules, and collections. By seeking to discover more about The Museum's activity through "data about data," our clients gave us a very compelling and forward-looking problem to investigate.

1. Data Overview

From the outset, much of our work involved processing the data down to a manageable size for visualization and identifying meaningful components and variables in the dataset that we could rely on to build our visualization. We initially managed the files sent to us by mounting them on Karst, a real-time, high-throughput computing cluster at Indiana University. In consultation with our clients, we proposed a dynamic animation built in D3.js that would track the scale and frequency of activity on database entities over a six-year period (2008-2014). The outcome is a web-based visualization that can identify what our clients characterized as "moving patterns of metadata around database objects" as well as a reproducible workflow that can be reapplied to the Museum's data.

Metadata from the EMu database system was provided to us in the form of Audit Records in CSV files associated with records and objects that are part of a new permanent exhibit at the Museum, the *Cyrus Tang Hall of China*. The primary data fields in the most low-density file consisted of IRNs (Individual Record Numbers related to objects, specimens, and entities), users (persons making changes), dates and timestamps (of changes), types of changes, and values changed. In addition, high-density files with 50+ columns included data identifying fields, departments (botany, anthropology, etc.), collections, provenance, and a host of other elements.

Following our first video conference with The Field Museum staff, we identified eight separate processes we determined would be useful to identify in the data.

1. changes recorded to IRNs/specimens
2. frequency of changes to IRNs, including burst analysis

3. total number of changes to specific objects
4. trend of changes over set temporal intervals (e.g., by month)
5. changes by metadata category
6. changes by users
7. changes by collection or type
8. the top N% of IRNs (i.e., objects or specimens) to include

An analysis of the primary metadata associated with the *Tang Hall of China* exhibit identified 20,059 total audit records of activity, with the following characteristics:

- **CatalogIRN:** 1818 unique IRNs. Across the time sample, the number of activities ranged from 2 to 89 for unique numbers. The average IRN had approximately 11 changes across time. The median of the number of changes per IRN was seven.

- **AudUser:** Database username identification. There were 100 unique user names. However, one of these is designated as emu, which is the administrative username for the EMu system. The emu ID alone has 8679 activities across the time sample of the data set, which accounts for 43% of the activity in the data set. In fact, across the time sample, four users — including emu — account for 80% of the activity within the data. Once the emu username was filtered, three users accounted for approximately 37% of the activity. Across users, activity ranged from 1 to 3833 operations. The median activity of users is eight operations.

- **AudDate:** Date of access and operation. The data sample ranges over six years of activity within the system. The time sample of the data occurs between 2/18/2008 to 9/5/2014. While operational activity remained fairly stable for the first few years, more activity has occurred within the last two. In fact, 2014 shows the highest number of operations.

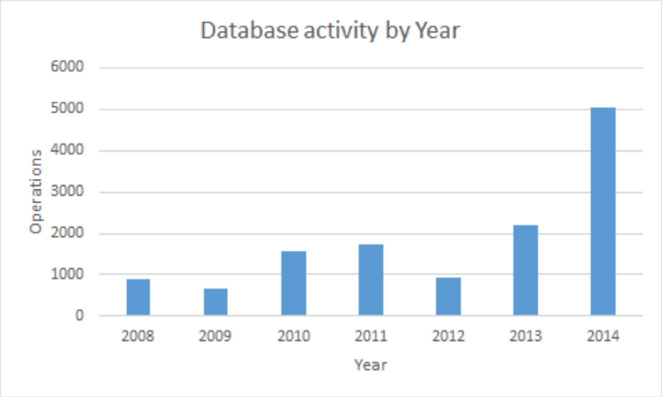


Figure 2: Database activity by Year (created with SAS)

- **AudTime:** Time of access/operation. From our point of view, the particular time of operation was not essential to understanding larger trends and patterns desired by our clients. Operations occur rather uniformly across time and at all hours of the day.

- **AudOperation:** Type of operation performed. Three types of operations are recorded: update, insert, and delete. Across the six years, the majority of operations, 12,632 occurrences, have been updates. There were a total of 4,520 delete operations and 2,907 insert operations across 6 years.

- **AudModule**: The database unit to which the information belongs. There were 16 different categories of metadata organization. The following table shows the distribution of activity within each module across the six-year sample data set.

Table 1

Module	Count
ecatalogue	12522
ecollectionevents	3623
econdition	186
edarwin	9
efmnh	2
efmnhtransactions	23
einternal	827
elocations	20
emultimedia	1193
eparties	105
eregistry	16
esites	330
estatistics	16
etaxonomy	601
etest	6
ethesaurus	580

2. Related Work

Visualizations of different information densities could be derived from our data. A high information density graph, for instance, would have represented every activity occurrence in the data. IBM researchers, for example, have created dense chromograms (fig. 3) to "analyze the huge edit histories" of Wikipedia administrators. The graph, they explain, "can display very long textual sequences through a simple color coding scheme" to "describe a set of characteristic editing patterns." [2]

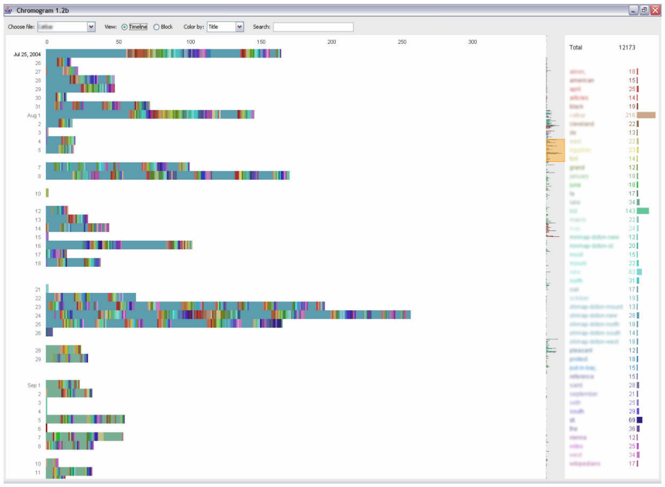


Figure 3: Wattenberg et al. 2007

Our early concept drawings (as in fig. 4) envisioned a high information density interactive visualization that would provide hover capabilities to view details about the record of each database transaction. After some deliberation, we chose to focus on producing a low-density, looping animation to better illustrate the metadata "flows" that interested our clients.

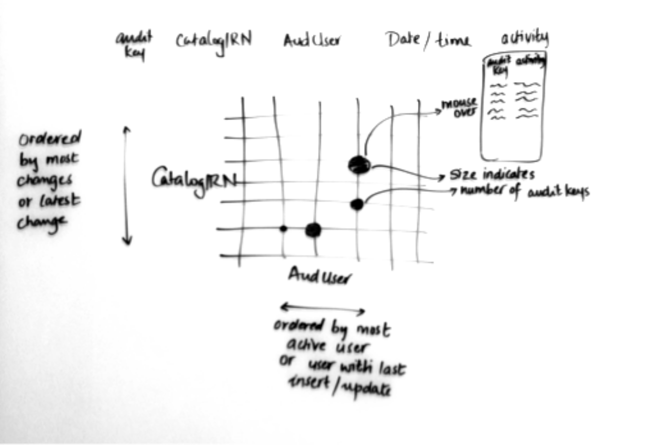

audit key
CatalogIRN
AudUser
Date / time
activity
audit key
activity
mouse over
Ordered by most changes or latest change
CatalogIRN
Size indicates number of audit keys
Aud User
ordered by most active user or user with last insert / update

Figure 4: An early whiteboard concept drawing

Another option would be covariance graph, in which the correlation between two key variables is illustrated. A series of such visualizations could also be useful for the aims of our project. An example of this type appears on The Field Museum's own website, where it is bundled with details about schemas and other aspects of the Museum's EMu implementation. According to the documentation accompanying the network graph, Fig. 5 was created using Gephi's ForceAtlas2 algorithm [3] and published online with the Sigma Javascript library.

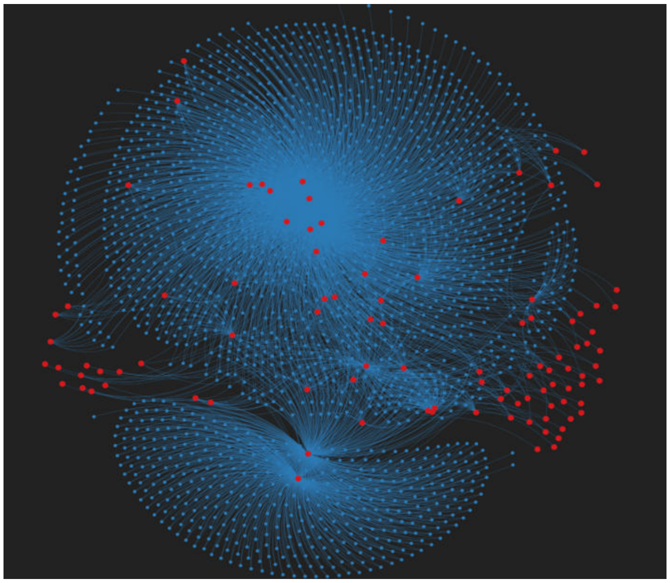


Figure 5: "Visualizing User Activity in EMu" http://emudata.fieldmuseum.org/emu-people-map/

"Visualizing User Activity in EMu" is, like ours, a way of using metadata to better "see" database activity (the network graph relies on IRN or "record" data, as we do). Where our application differs is in our aim to a temporal dimension. Time-ordered network analysis exceeds our present skillset. Moreover, although the data used in fig. 5 is identified with the China Hall records, just like ours, the focus is specifically on "multimedia records" (from the "emultimedia" module). Instead we made the decision to rely on "catalog records," or those from the "ecatalogue" module, because of their higher occurrence rate in the EMu dataset we received. In the audit records we analyzed, the AudModule field "emultimedia" appears only 1,193 times, whereas the field "ecatalogue" occurs 12,522 times. Nevertheless, the workflow we have produced for ecatalogue records can also be applied to emultimedia records and their relation to other database objects.

3. Analysis

The results of our analysis and validation processes were profound in many cases. We made a number of important adjustments to what ultimately became our visualization dataset as we learned more about the data. Among them,

- revised our approach to reflect new data from the client
- shifted focus to a single module — "ecatalogue"
- connected audit records to specimens and objects (and thus Fields and Departments) with unique catalog IRNs
- reduced the data to 19,178 observations
- reduced the role of "users" as a data variable

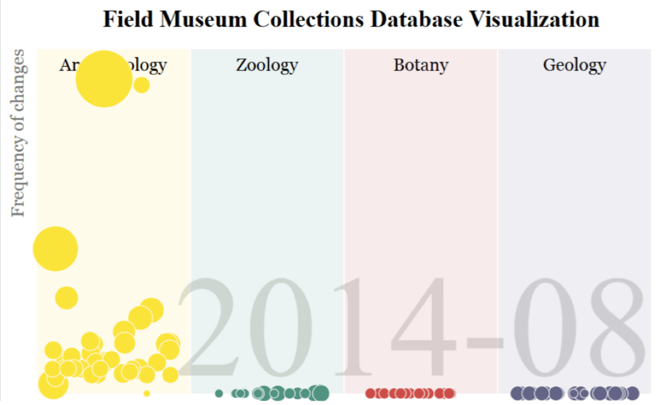


Figure 6: Screen grab of EMuView in its first iteration.

Because IRNs are not unique across different components and modules of the EMu database (i.e., emultimedia IRNs could be identical to ecatalogue IRNs, but refer to different actual objects) we eventually included only the largest module type, ecatalogue. This enabled us to join the catalog table provided to us — and hence field and department variables — with the metadata audit records using the unique catalog IRN. We also filtered out an administrative user in the metadata ("emu"), which alone represented 43% of activity related to merely routine maintenance tasks. We determined that time of day and type of action were not relevant to account for in the visualization, so we removed those fields as well from our data. Finally, we sorted out a top-N sample of approximately 2.5%. We originally began with a sample of 2%, which when aggregated on IRNs came to 6,117 row. But this initial sort resulted in only two departments being represented, Anthropology and Zoology-Insects.

Table 2

Field/Department	Total number
Anthropology	11478
Botany	2191
Geology_Fossil Invertebrates	92
Geology_Paleobotany	3285
Zoology_Amphibians and Reptiles	243
Zoology_Birds	255
Zoology_Fishes	115
Zoology_Insects	1111
Zoology_Invertebrates	408

Consequently, we added highest frequency entities from all fields and departments based on measuring the proportionality of the departments against the sum of all the database entities.

The animation loop that we developed in D3.js relies on JSON data input, which contains the following variables:

- **circles**: individual entities identified by catalog IRNs
- **vertical bands**: each of the four fields at the Museum
- **up and down movement** (y-axis): database activity by entity per month, with new circles spawned as the appear in the dataset
- **growth of circles** (x-axis): cumulative changes made to entities over six years, enlarging in proportion to the cumulative sum of changes for each
- **motion duration** (z-axis): captured in monthly intervals

The Museum staff endorsed our vision for a dynamic rather than interactive visualization, although they also suggested that the loop be able to pause with a keyboard stroke.

A demonstration of the animation can be seen at http://ella.ils.indiana.edu/~kadur/IVMOOCFinal/.

4. Workflow

We relied on the usual tools for working with CSV files, such as Excel and various text editors. The core of our workflow consisted of data processing with RStudio (and RStudio Server running on the Karst cluster), with help from SAS for visual analysis and our relative frequency calculation.

A complete set of our code, input data, and workflow can be found at http://ella.ils.indiana.edu/~rshiggins/ivmooc. A few of our operations were fundamental to the success of our work, which for convenience we've reproduced, as follows:

Filter and Merge (R script)

```
library(dyplr)
eaudit01 <- read.csv(file="MOOCeaudit03.csv",
       head=TRUE, sep=",")
  eaudit02 <- eaudit01[,c(2, 3, 4, 6, 7, 8, 11)]
  eaudit03 <- eaudit02[eaudit02$AudUser != "emu", ]

catalog <- read.csv(file="ChinaHall_masterRev.csv",
       head=TRUE, sep=",")
  aucatalog <- merge(eaudit03, catalog)
  aucatalogVars <- aucatalog[,c(1, 2, 3, 4, 5, 6,
       8, 9, 11, 26, 38, 39, 40, 41, 42, 43, 44, 6
       0)]
```

Cumulative Counts (R script)

```
cumuldata <- sapply(1:length(data$CAT_irn)
       ,function(i)
  sum(data$CAT_irn[i]==data$CAT_irn[1:i]))
  cumulfile <- cbind(base, cumuldata)
```

Relative Frequency by Month (SQL Query in SAS)

```
create table max as
  select cat_irn, year, month, max(cumulative)
  from freq
  group by cat_irn, year, month
```

Top(100) Filter (R script)

```
topN <- tbl_df(data) %>%
  group_by(date) %>%
  count(CAT_irn) %>%
  top_n(100)
```

All pre-processing and calculations were completed in RStudio to minimize the load on our front-end web-based scripts. Once we tailored the data down to specific variables, the CSV data was converted to JSON for ingestion into the D3 code. The D3 framework is embedded in the standard components for a website, consisting of an HMLT page, Javascript files, including JQuery, and a CSS folder.

The D3 code itself is a modified version of Mike Bostock's recreation of a visualization entitled "The Wealth & Health of Nations," which was originally developed by Gapminder and used by Hans Rosling for a 2006 TED talk. Tom Carden also contributed code to Bostock's example. [3]

5. Challenges and Opportunities

One important challenge we faced was processing the data into a manageable form for a visualization. We spent considerable time analyzing the components of the database and testing scripts that would process the data effectively and accurately. We also sought a unifying unique ID that could provide a fixed point over which our representation of the metadata could flow. After realizing that IRNs could appear in approximately three different forms (roughly, Catalog, Module, and Record IRNs) that were not unique to each other, we focused on the module that was most closely associated with catalog IRN. With more time, we may have been able to parse multi-value "tab" fields in the audit records in order to include additional modules.

Another problem we had to solve occurred when we began working on the front-end of our project. We discovered two values would be necessary to embed our data into a temporal frame consisting of intervals: cumulative frequency and relative frequency. The first was the growth of each entity, which could be accounted for with accumulating sums for each individual occurrence of activity. But we also had to account for up-and-down movement, or the extent of the activity, which we eventually solved by calculating relative frequencies for each of the catalog objects per month. Cumulative counts capture the activity over the entire period, whereas relative frequency ratios are currently calculated by month.

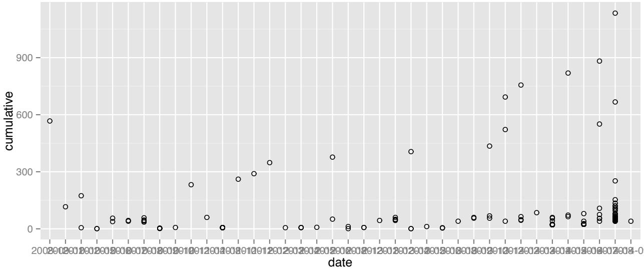


Figure 7: Scatter plot showing sharp increase of activity over time

Before fully completing the visualization, we have plans to test the D3 code with larger sample sizes by doubling and tripling the amount of entities ingested into the algorithm. We will select additional entities based on sorting by high relative frequency ratios to add more activity to the overall animation. Additionally, in the near term, more attention could be given to the D3 code base for the project. Tweaking the "transition" function, adding "ease" to the dynamic of the movement, and integrating more context to the movement across the loop would make the visualization more effective. Placing a static plot diagram below the loop or adding tracing trails to the circles' trajectory are some of the options that would increase the visualization's cognitive impact. Moreover, in addition to the pause/play button in our design, the incorporation of an interactive "slider" just below the animation frame would allow users to have complete control of when and where they wanted to freeze the graph.

Going forward, we would also be interested in seeing an extended representation of all 20,000+ audit records in the China Hall dataset, including those in different modules. For example, finding the right way to associate modules such as *ecatalogue* and *emultimedia* would likely not be unreasonably difficult, although we did not have the time in this round to make these connections.

6. Conclusion

Some basic facts about the China Hall data emerged in our work. Anthropology is by far the most active department, especially after 2013. Zoology, Botany, and Geology show occasional bursts of activities during certain periods of the year, and this raises questions regarded the context for those bursts. Most of the recorded activity rises in volume leading up to 2014 (see fig. 7), with very little activity in 2008. This would seem to be consistent with the coordination of the China Hall material for the opening of a permanent exhibit. We wonder what differing patterns of activity and departments would look like in records selected from different collections or in a cross-section of records from a specific period.

Now that a workflow has been designed for this process, we would be especially interested in seeing larger portions of EMu records measured and visualized with these processes. Applying some version of our workflow to the 24 million specimen records in the Field Museum's EMu system, as well as the roughly 53 million audit records collected in the database, is a challenge we hope to see undertaken in the future.

7. Acknowledgements

We are very grateful for the engagement, interest, and feedback of our clients in this project, Sharon Grant, Katherine Webbink, and Marc Lambruschi. We gained invaluable experience thanks to their generosity. And we would like to thank our teachers, Prof. Katy Börner and Michael Ginda, for the inspiration and insight they gave us during the development of our work.

8. References

[1] S. Grant. Drupal or bust: The hazards of mixing technologies. Slide Presentation. KE Software company website, 2011. https://emu.kesoftware.com/downloads/EMu/UserGroupMeetings/2011_Global/pps/sgrantemu2011.pps

[2] M. Wattenberg, F. Viégas, and K. Hollenbach. Visualizing activity on Wikipedia with chromograms. INTERACT, Part II, pages 272 – 287, 2007.

[3] M. Bostock. The Wealth & Health of Nations. D3.js. 2012. http://bost.ocks.org/mike/nations/; Gapminder. Wealth & Health of Nations. 2010. http://www.gapminder.org/world/#;example=75; H. Rosling. The bests stats you've ever seen. Ted Talk. ted.com, 2006.; T. Carden. MindGapper. 2011. http://randometc.github.io/mind-gapper-js/