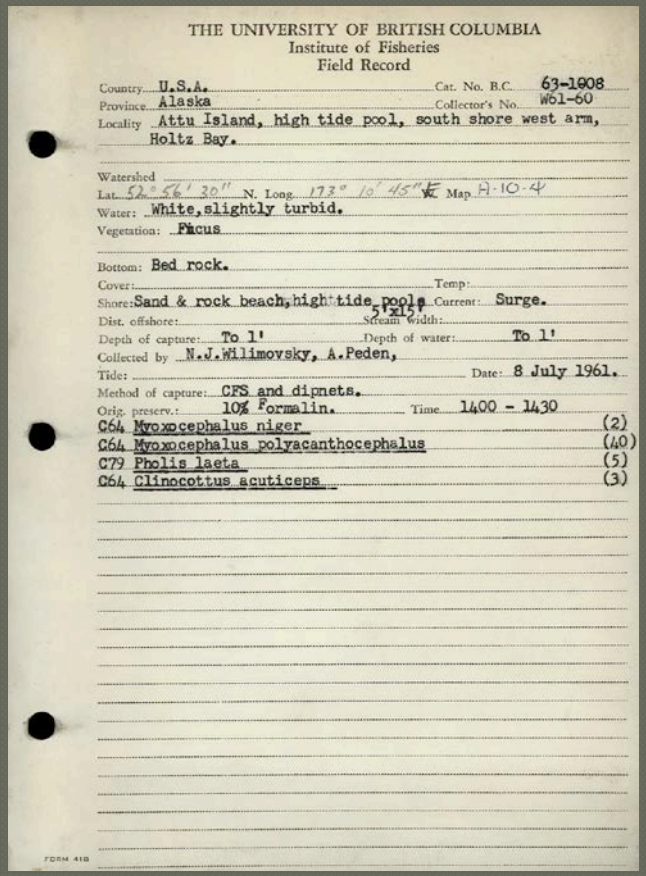
The UBC Library Digitization Centre

Making Big Data Available: The University of British Columbia Institute of Fisheries Field Record Lab Notebooks Experience

Robert Stibravy, Digital Projects Librarian

The Problem

- What do you do when you have 11035 pages of this…



THE UNIVERSITY OF BRITISH COLUMBIA
Institute of Fisheries
Field Record

Country U.S.A. Cat. No. B.C. 63-1908
Province Alaska Collector's No. W61-60
Locality Attu Island, high tide pool, south shore west arm, Holtz Bay.

Watershed
Lat. 52° 56' 30" N. Long. 173° 10' 45" W Map A-10-4
Water: White, slightly turbid.
Vegetation: Fucus

Bottom: Bed rock.
Cover: Temp:
Shore: Sand & rock beach, high tide pools Current: Surge.
Dist. offshore: Stream Width:
Depth of capture: To 1' Depth of water: To 1'
Collected by N.J.Wilimovsky, A.Peden,
Tide: Date: 8 July 1961.
Method of capture: CFS and dipnets.
Orig. preserv.: 10% Formalin. Time 1400 - 1430

C64 Myoxocephalus niger (2)
C64 Myoxocephalus polyacanthocephalus (40)
C79 Pholis laeta (5)
C64 Clinocottus acuticeps (3)

FORM 418

The Problem (cont.)

- …and you want to find all 311 occurrences of this species in the UBC Institute of Fisheries Field Record collection?

Oligocottus maculosus

- Or some other salient data from the collection?
- You could…

A "Solution"

- Call Rick Taylor at UBC and get him to…
  - Look for the data
  - Photocopy the relevant pages
  - Mail or fax or email those pages to you
- Let us just hope that you only have one of these data requests in your career! (Rick is a nice guy, but come on!)
- Or…

The Solution!

- Digitize the collection and make it full-text searchable
- Put it in a content management system (CMS) and allow (encourage!) global access and linking to the data (and the original images of the notebooks)
- Hook the data in to the FishBase database so that users will see these in their search results

The Work

- Digitize all 11035 pages
- Post-process these to correct for brightness, contrast, sharpness, etc.
- Optical character recognition (OCR) for the parts of each notebook page that are amenable to OCR, transcribe the rest by hand
- Load in to the CMS
- Export metadata to allow linking with FishBase

Fujitsu fi-6670A

High speed document scanner – this is what we scan large document collections with

The Work (cont.)

- The scanning and the post-processing go well – we have done this before
- But (there is always a "but") the OCR is a washout – and that leaves us with a single choice
- Full Hand Transcription!
- Uh oh!

The (New) Solution

- Hand transcription – at first look this does not seem to be practical
- Sampling various pages, however, and testing these with several different people doing the transcription gives us a result of 12-15 months with two students working on it and my project management
- In the end it takes about 14 months

Results

- 11035 pages of difficult-to-access data available to the world
- A proof of concept for future endeavours of this nature and the realization that "it can be done!"
- Cost: approximately $22K CAD (not including existing infrastructure)
- Nice to have: domain expertise for better quality control and possibly faster work

Quality Control (QC)

- With a large, hand-transcribed project QC is essential to ensure a high degree of accuracy
- This is particularly true when dealing with scientific data that will be used for research and policy development that effect lives and livelihoods
- …and then there is the QC *after* the QC (well, a different kind)

Quality Control (continued)

- Because these data will be part of FishBase, the largest fish and fisheries database in the world, normalization is required before the data that is currently in CONTENTdm can be exported for use
- We are working with the lead FishBase programmer with a target of either this month or March 2015 for the first export of the data

Next Steps

- Once the FishBase work is done we will be adding these data to the Global Biodiversity Information Facility (GBIF)
- GBIF is an open access, common standards data repository covering all types of life on earth
- We have asked our project sponsor, Dr. Rick Taylor, or his colleagues to consider other potential projects with us

Q & A / Thank You

- Questions or Comments?
- Thank you/谢谢/Terima kasih/有り難う/多謝/감사합니다/Whakawhetai ki a koe/Ngeniyerriya/Mahalo/Obrigado!