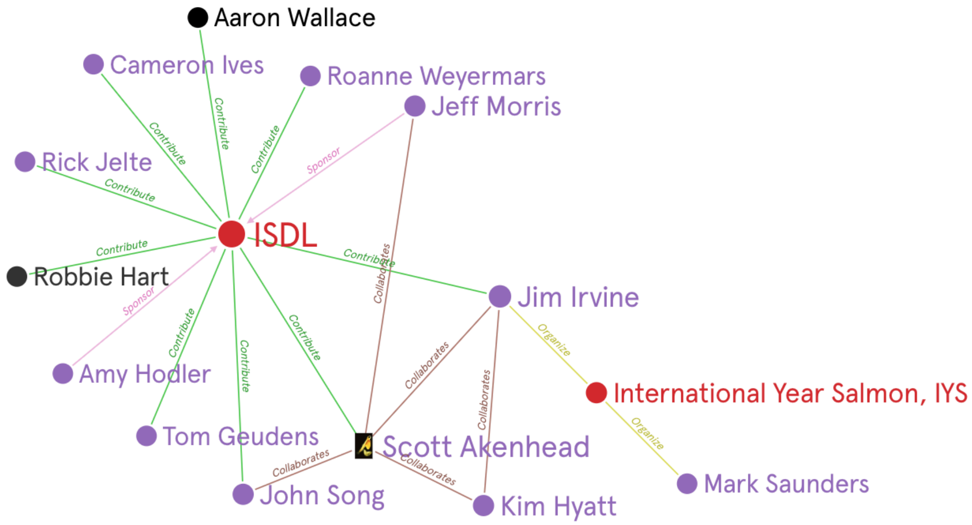
A Prolegomena[1] for ISDL

The International Salmon Data Laboratory (ISDL) will deliver examples of modernizing information flows in salmon research and management.

Date: 2018-10-5



Authors, not ordered

Name	Organization	Contact	Roles
[your name here]			
Scott A. Akenhead https://orcid.org/0000-0003-1218-3118	Pacific Biological Station Fisheries and Oceans Canada (DFO)	scott@s4s.com 1.250.210.4410	instigator of ISDL; analyst
James R. Irvine	Pacific Biological Station Fisheries and Oceans Canada (DFO)	james.irvine@dfo-mpo.gc.ca 1.250.619.8197	salmon ecologist; Canadian lead: International Year of the Salmon
Kim Hyatt	Pacific Biological Station Fisheries and Oceans Canada (DFO)	kim.hyatt@dfo-mpo.gc.ca	salmon ecologist
Tom Geudens	Neo4j Corporation, Halle, Belgium	tom.geudens	Neo4j staff: field team member
Jeffrey M. Morris	Neo4j, Inc., San Mateo CA	jeff@neo4j.com	instigator of ISDL; sponsor

[1] https://en.wikipedia.org/wiki/Prolegomena_to_Any_Future_Metaphysics

John Song	Systum Inc., Los Angeles CA;
	Ho Chi Minh City, Vietnam;
	Ireland; UK; Plano TX;
		jsong222@gmail.com
		1.323.823.6700
			graph DB expert; “in pursuit of hidden discoveries that further expand the collective knowledge of processes that govern our world”

Richard Jelte		imjelte@gmail.com	
Cameron Ives		cives@cellsignal.com	
Roanne Weyermars		roannebw@gmail.com	
Amy Hodler	Neo4j, Inc., San Mateo CA	amy.hodler@squareyolk.com	
Robbie Harding		robbieh@stanford.edu	
Aaron Wallace		Aaron.Wallace@pb.com	
Dillon L. Shook		dshook@alumni.nmt.edu	
Bruce Hecht	Analog.com	Bruce.Hecht@Analog.com	
Guillaume Pharand		gpharand@gmail.com	
Semih Salihoglu	University of Waterloo, Waterloo ON, Canada	semih.salihoglu@uwaterloo.ca	Graph DB expert, academic
Mark Saunders	International Year of the Salmon	msaunders@yearofthesalmon.org	sponsor
Rick Jelte	Ultimate Software	imjelte@gmail.com	UI/Javascript (front-end) Software Engineer Graph visualization enthusiast. Level of Salmon expertise: Novice

Table of Contents

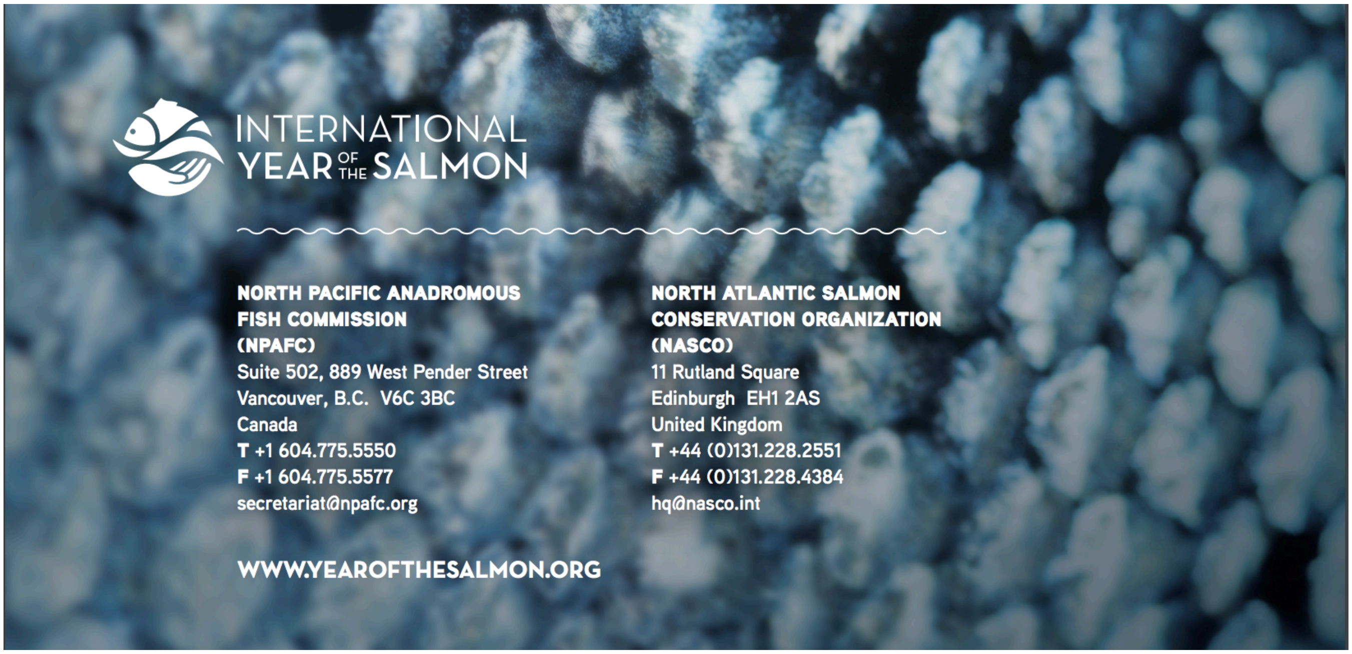
Background

Salmon

The Wikipedia article on salmon is generally excellent[2] if occasionally "enthusiastic" and touches on many of the issues facing salmon.

International Year of the Salmon (IYS)

ISDL is an activity within IYS. See descriptions if IYS on websites for North Pacific Anadromous Fish Commission and North Atlantic Salmon Conservation Organization.



INTERNATIONAL
YEAR OF THE SALMON

NORTH PACIFIC ANADROMOUS FISH COMMISSION (NPAFC)
Suite 502, 889 West Pender Street
Vancouver, B.C. V6C 3BC
Canada
T +1 604.775.5550
F +1 604.775.5577
secretariat@npafc.org

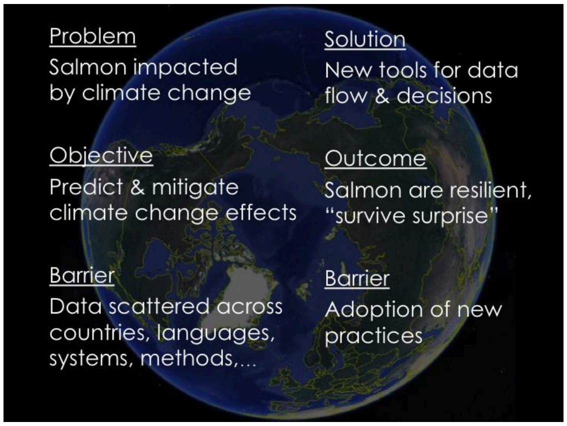
NORTH ATLANTIC SALMON CONSERVATION ORGANIZATION (NASCO)
11 Rutland Square
Edinburgh EH1 2AS
United Kingdom
T +44 (0)131.228.2551
F +44 (0)131.228.4384
hq@nasco.int

WWW.YEAROFTHESALMON.ORG

[2] This article would be better if more salmon biologists contributed.

ISDL at GraphConnect

ISDL was announced as a Graphs For Good initiative supported by Neo4j Inc. at GraphConnect, NYC, 2018-09-20. See the 18 minute video of the presentation by Scott Akenhead. Slides on same link.



Problem
Salmon impacted by climate change

Solution
New tools for data flow & decisions

Objective
Predict & mitigate climate change effects

Outcome
Salmon are resilient, "survive surprise"

Barrier
Data scattered across countries, languages, systems, methods,...

Barrier
Adoption of new practices

https://yearofthesalmon.org/project/international-salmon-data-laboratory/

Abstract:

The International Year of the Salmon (IYS, https://npafc.org/iys/) involves a dozen countries with one overarching goal: **salmon are resilient to climate change**. Survive the surprise. Salmon uniquely integrate terrestrial and marine effects from global warming. They are the canary in the coal mine. Help salmon help us. This requires (a) radically better information flow and knowledge management—new best practices; and (b) data mobilization, so we actually know what we should know from the data and metadata we collected—graphite to silicon.

Salmon biologists need new tools for data assembly, analysis, and visualization; tools that are powerful but safe, effective but easy-to-use, and rewarding. Hey, how hard can that be? Hard enough to need help from new friends. We are launching the International Salmon Data Laboratory (ISDL) so that you can help us leverage the power of Neo4j and new tools built on neo4j: Morpheus, Bloom,. The tools you are developing can have impacts throughout the environmental sciences. The revolution begins with salmon.

From two years of IYS preparation, we have plans, participants, and vision for the ISDL. We need technology, technology support, and the ability to support participants and users. We need funding to keep the ISDL glued together long enough to deliver something truly transformational. And to get those tools back to participants.

Vision, Goal, Objectives

Goal — *the revolution begins with salmon.*

By 2022, examples of how new IT has solved a series of problems related to information flow in salmon research, management, and communication have resulted in wide adoption of that IT by salmon biologists and initial adoption in other sectors of environmental science and management.

Objectives

Datasets
Salmon datasets are, or will be, posted on-line as the basis for examples of data management, analysis, and communication. How any graph databases built from contributed datasets will be managed is to be determined (TBD).

Reusable examples
Demonstrations of what can be done using new tools (available before 2022). The intention is that biologists who are not IT experts can re-use and re-purpose the examples in their situation. The examples should be sufficiently easy-to-use, accessible, and rewarding that biologists and managers will adopt new practices.

Sandbox
A server holds all related datasets, software, code, tools, examples, results, and reports.

Software
Licenses for commercial software are available for ISDL examples and demonstrations.

Volunteers
Salmon experts and graph-database experts work together to:
(a) identify and describe worthwhile and instructive problems regarding salmon data;
(b) provide a link to the dataset representative of each problem;
(c) solve those problems by applying neo4j database technology and related tools including workflow software and analytical software (R, ML, AI)
(d) prepare examples, based on the provided datasets, for participants (new and seasoned) to discover and explore.
(e) repare brief technical reports valuable for encouraging other experts (both types) to participate in ISDL and to apply the examples in their situation (SW licences may be required).
(f) assist novice users to understand and reapply the examples.

General Problems

Metadata

In a time-series of salmon data, every data point (every observation) has a story. Without those stories, the analyst may be misled, thence the decision-maker. Metadata improves analyses by identifying (a) factors that need to be included (else missing variables), and (b) the precision of observations, allowing poor observations to be down-weighted (have less effect on results) or excluded.

Credit: Chris McLennan, intrepid.photos

raw data

Details like sample size, timing, location, method, and data processing in the field may vary among observations in raw data. ***Example*:** where sockeye salmon smolts are emigrating from a lake to the ocean, important data is obtained when there is a fence across a river to direct smolt to where they can be counted by photographs or in traps. Logistics problems[3] can mean a fence was deployed later in the smolt emigration than intended or removed earlier. What data was actually obtained? What was done to "repair" the resulting estimate? Is there a better way to accommodate truncated sampling when estimating total abundance (*hint*: analyze years together instead of individually)?

Old data collections designed for relative abundance estimates but otherwise uncalibrated, may be compared to more recent data collections that have been calibrated. This can cause a false alarm.

aggregated data

The precision and reliability of each point is worth knowing before interpretation and analysis. ***Example**:* in the NE Pacific Ocean a complicated "mixed fishery" has catches from all of the populations of Fraser River sockeye salmon. Some of the catches are sampled for population[4] composition by genetic ID (. Mixture analysis is applied to the catches and samples to generate catch by population Errors in population identity mean the catch estimate for small population tends to be overestimated. ***Example**:* 2 populations in catch, one is 95% of catch, classification error is 1%. The smaller population is overestimated by 18%.

analyzed data

Analysis amplifies the importance of precision and reliability in the original data. ***Example**:* salmon productivity is typically characterized by returns/spawner or *R/S*. The probability density distribution (PDD) of a ratio is lognormal if the components have a Gaussian (normal) PDD. When productivity is expressed as log(*R/S*) you might expect that estimate to have a Gaussian PDD. But estimates of *R* and *S* are imprecise when relatively small (variance is not stationary). Severe underestimates of *S* can produce extreme overestimates for *R/S* that are outliers even as log(*R/S*). That possibility is an issue if you estimate the Ricker curve via

[3] Field notes for Hobiton Creek (NitNat Lake) counting fence 2016: "Bear demolished fence July 9." Field notes for Tahsis River survey: "In 2012 and 2013 there were several run-ins with belligerent bruins that were in the channel feeding on spawners. On two occasions the bears refused to move out of the channel and we had to start the survey at that point. Bear spray is required equipment."

[4] In practice, closely-related populations are grouped into *conservation units* (CU, a.k.a. stock, a.k.a. metapopulation). CUs are further grouped into *runs* based on timing of return to the coast (river entry) from the open ocean. Management has a trend to increasingly fine-grained. We can now a salmon's DNA to determine the spawning site (population, sort of) of that salmon's parents, via DNA samples taken previously at a that spawning site.

$$\log\left(\frac{R}{S}\right) \sim \alpha - \beta S + \mathrm{N}(0, \sigma)$$

because the outlier at low *S* will obviously cause **α** (the Y axis intercept) to be overestimated, but that means **β** is also overestimated. That combination indicates a highly productive population with maximum sustainable yield at a low stock size (S_{MSY}). From Hilton (1997),

$$S_{MSY} = \frac{\alpha}{\beta}(0.5 - 0.07\alpha)$$

"Fish hard" is bad advice if you are uncertain.

Fixing the metadata problem

Understand the data or risk that your analysis will be misleading (garbage in, highly sophisticated garbage out). This will involve:
(a) discovering and mobilizing the appropriate metadata, simply more data;
(b) linking that metadata to the core data (perhaps within a knowledge graph, but as points or as datasets?); and
(c) taking full advantage of the additional information in the metadata during analyses[5].

Links between datasets

Certain metadata allow links between datasets, perhaps to allow abundance estimates for a salmon life stage at some place and time to be matched with corresponding habitat indicators. *Example:* egg to fry survival is affected by river flow velocity (high: scouring, low: desiccation) and temperature (freezing, development rate) at the spawning site. Matching hydrology data to estimate of spawner and fry abundance (or any later life stage) requires (in USA and Canada) identifying a hierarchical watershed code (or name) to the name (or coordinates) of the spawning site.

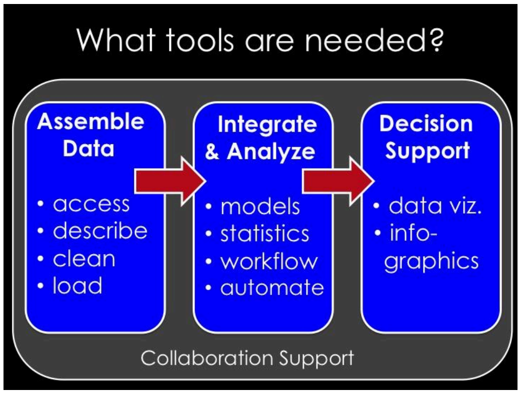
Disparate datasets

Add More

[5] At the risk of self-aggrandizement, this paper applies metadata for relative precision (uncalibrated but ordinal) and previously unrecognized events (a spawning channel that kills fry) in an analysis of salmon productivity:
Akenhead, S.A., J.R. Irvine, K.D. Hyatt, S.C. Johnson, C.G.J. Michielsens, and S.C.H. Grant. 2016. *Habitat manipulations confound the interpretation of sockeye salmon recruitment patterns at Chilko Lake, British Columbia.* NAPFC Bulletin 6: 391−414.

Technical Vision

Here is a conceptual description of a desirable salmon information flow, from field to decision. Please add examples of software, process, data, etc.



What tools are needed?

Assemble Data
- access
- describe
- clean
- load

Integrate & Analyze
- models
- statistics
- workflow
- automate

Decision Support
- data viz.
- info-graphics

Collaboration Support

1. **Raw data**
   Raw data from field and laboratory operations is digital (not on paper) and subject to quality assurance (QA) checks during the collection process. This data is immediately uploaded to the Web.
    a. Data examples:
        i. stream survey to estimate spawner abundance.
    b. Tool examples:
        i. GeoViz by Sierra Technologies.
        ii. 703,000,000 hits from Google search: “mobile forms.”
        iii. See: Google Forms, Device Magic, Nexticy, iFormBuilder, Formhub, TrackVia, Magpi,.
2. **Field data**
   Upon upload, field data is subject to quality control (QC) checks beyond what is practical in the field. QC results are sent to collector. As a result, the dataset may be replaced or deleted.
    a. Data examples:
        i. Data from a survey of a stream are examined (QC), plotted, and returned in the context of previous and related surveys. Computations such as “area under the curve” and extrapolating patterns from preceding and related surveys.
        ii. Consider that data for a recurring survey can accumulate in a file as .csv, .json, or SQLite (a relational database stored as a file instead of mapped to sectors of a hard drive), and referenced from a node in a neo4j knowledge graph. Alternatively all pieces of data are named fields in nodes linked within and between surveys.
    b. Tool examples:
        i. R, Shiny, Rserver.
    c. Process examples:
        i. Taverna workflow will trigger R scripts via Rserver to produce on-line, interactive graphics via Shiny. There are many similar toolsets for this, but R is widely used by ecologists.
    d. Result examples:
3. **Integration**
   Data is only meaningful in the context of other data. That can be obtained through a large knowledge graph as the context for analysis, reports, and interpretation.
    a. Data examples:
        i. Spectacular lack of context: 42 (HHGTTG)
        ii. After a new field observation is immediately uploaded, the collector receives a report putting that into the context of preceding and related observations.
        iii. [add more]

4. **Analysis**
    a. Example
        i. Extrapolation of parameter estimates from data-rich population to data-poor. This is important for categorizing the status of salmon populations which involves using estimates of habitat availability to guide extrapolation of parameter estimates[6].

5. **Products**

6. **Communication**

ISDL Architecture

“To begin leveraging Neo4j for analyzing the salmon datasets, a core data schema will have to be established. Graph Commons provides a tool for collaborating on the schema design. Using the Darwin Core model to drive the initial design would be a good starting point. Once the core schema is generated, creating publishing pipelines into the schema from disparate datasets would be the next natural course of action. The Neo4j schema is forgiving and extensible, so the data model can evolve with deeper insights into the available datasets.”
“I've created a schema graph at **graphcommons.com**. If you search for "Salmon Data Laboratory", it will pop up the canvas as a "Work in Progress". It is a public graph. Although the tool is for visualizing real data, it can be used to visualize a data schema model. There may be better public tools out there, but this seems to have all the features needed for collaborating on a public graph.” - @JS1

Size

Q.: Size and growth prediction for the data? **A.:** The largest dataset is NuSeds, a table of about 390,000 rows of ~30 variables. We are likely to work with a small subset of NuSeds but needs to be linked to extensive detail (in the graph domain). I forsee max 10 datasets for ISDL examples in 2018, max 100 by 2022. The problem is not size, it is assembly (access), standardizing (ontology required), cleaning (synonyms, classification of practice, etc), integrating, analysis, products, and the automation of all these. I will post links to data at community.neo4j.com as it tumbles in.

Public facing?

Q.: If so, in what way: application, Neo4j browser connection,? **A.:** Yes. I want many people need to see the examples. Maybe webpages via Structr, or something simpler to start. It should be possible for people to look closely (Bloom, Browser) without being able to break anything.
I think 200 users would be a howling success. 2,000 means we fell through a wormhole.

Access

[6] For example, treating habitat as nursery lake area per salmon to guide estimation of a population “capacity” parameter for conservation units within the Fraser River sockeye salmon:

Q.: Will some persons have different access the database (i.e. Bloom)? Is that a separate design requirement?
A.: Inevitable. Developers have more permission than known users (login) have more permission than public.

Security

Q.: How do you see security? **A.:** Lax. Users (login/pw) will generally be previously known, and we simply boot malfeasants. The data is of value to a small community. Providers can have a copyright, and attribution is a critical aspect of sharing.
Please steal the examples! To re-apply them in an institutional/enterprise setting, SW licenses may be required.

Query Load

Scant. Certainly to start. In the happy event that I am wrong, use will ramping up slowly and we will react.

Uploads

Q.: How do you envision uploading of data? Small transactional size datapoints, gigantic bulk sets of datapoints, or in between? **A.:** The data will arrive as .csv tables. ISDL is about tools. Other, bewildering, people want to maintain a jumble of jumbo datasets.

Verification

Q.: Should data contributions be verified first? **A.:** Yes, but no. I cannot imagine a user stuffing illicit or broken data into an existing knowledge graph as a covert act. Working with separated examples, that are backed up will minimize the effect of accidents.

Backup

Backup and roll-back might not be possible. BUT to re-use the examples in a new context implies that we maintain scripts or workflows (better?) that will recreate the example from scratch. Perhaps that means starting with workflows.

Frequency

Q.: How big will the daily updates/contributions be? **A:** Frequency: 0.1/day; Update size: microscopic; Contribution size: typically < 1M.

Database Design

To build a knowledge graph for salmon ecology, including related science and management, we need to define a rich set of ~15 **resources** (node types), e.g. `(:Person), (:Place)`. We will need ≤ 15 types within each resource. There will be many types of links between resources, e.g. `(:Person)-[HasPlace]-(:Place), (:Person)-[HasPerson]-(:Person),` and `(:Place)-[HasPlace]-(:Place)`. A link need a type to be informative, so that an "edge" (a triplet, node-link-node, see RDF) is maximally informative. Nodes are information, but the description of exactly how nodes are linked is knowledge, hence "knowledge graph." Resource types and link types will be the labels for graph.

```
(:Person{name:"Scott Akenhead"})
    -[HasPerson{label:"collaborates"]-(:Person{name:"Jim Irvine"})
```

This would display as Scott Akenhead - collaborates - Jim Irvine.

need details.

Specific Projects / Irresistible Examples

"In terms of how to approach the salmon data problem, I would suggest we start by modeling the data in neo4j. We can begin with the Darwin core model as the foundation of the graph and expand out the schema with further study of the different datasets available. Establishing the core schema will generate a target for mapping the different datasets into the core model which will drive analyses. Once that model is created, we can begin to build data publishing pipe lines that will read the data, validate the data, cleanse the data, and finally transform and load the data into the Neo4j graph. We can further attack the disparate dataset problem by building out web based and mobile tools for collecting data so that we solve the data problem at the collection level. Finally, we can perform analytics on the data, leveraging available tools such as Bloom or Tableau. Please regard the information above for what it is: an abstract plan of attack. There are many steps in between, but that should serve to guide the effort in a sensical path." JS 2018-10-3

Potential projects:

- Darwin Core
- Pipelines
- Web-based mobile tools for collecting data
- Analytics

Project 1. Standards

from Matt Denniston (thanks, Matt!):
http://www.streamnet.org/wp-content/uploads/2017/06/CoordinatedAssessmentsDES20170701.doc

The problems of standards and integration were not emphasized in the IYS Salmon Status and Trends (IYS-SST) Workshop #1 Vancouver 2019-01-23/24. But inescapable, lest oranges be taken for apples.
For example, here is the definition of SAR from StreamNet (above):

SAR	The point estimate for smolt-to-adult return rate, calculated as 100 X the point estimate of the number of returning natural origin adults, divided by the point estimate of the number of smolts that produced those returning adults.	Single	Required if NullRecord = "No”. Express these values as percentages (numbers from zero to one hundred), with two digits to the right of the decimal point. Examples: .020 = 2.00, .0015 = 0.15. This field holds a numeric value only -- the percent sign is implied but not included. Do NOT include repeat spawners in the number of adult returns. (A fish only returns once from smolting; subsequent returns are not appropriate for inclusion in smolt-to-adult estimates because they head to sea as adults on subsequent trips and thus are not exposed to the same suite of mortality factors.)

There are additional and related fields such as SARLowerLimit and SARUpperLimit. We did not thoroughly discuss the relative precision of estimates that need to be provided when data is shared. That was relegated to provision of metadata, yet to be specified. Look at this glossary’s definition of *Protocol and Method Documentation* as an example of what will be needed (and expanded) to capture just one aspect of metadata.

We have to decide if IYS-SST is a meeting for co-authors of a few descriptive papers, or the foundation of a much larger initiative that will involve many more datasets, ecologists, technologists, and years. If the former, I’m outa here. If the latter, we have to confront standardization. As in, whose standards (Darwin Core, Pacific States Marine Fisheries Commission, etc.)? What happens to our extensions of those standards? IYS-SST has broad scope, such as including multiple languages (the “I" in IYS), new derived terms (a Canadian CU is not quite an American ESU and I have no idea about equivalents in Russia, EU, Norway, UK,), and new methods with pointers to new documents. Exactly how do Russian biologists count spawners? Estimate total spawners?

I see this is an opportunity for new tools. Specifically, all of these terms and the cross-references to other terms within their definitions, can be approached as a graph. For example: The estimate of SAR is a numeric “indicator” within all of information about what SAR estimates refer to. SAR is an Indicator, Indicators are a subset of the SAR record definition.

I suggest we push all of this stuff about definitions for names into a sub-graph of linked ***Ideas.*** That way, a numeric estimate for SAR is just linked to the appropriate Idea nodes … `(Idea {name:’SAR’, … })`

(Idea {name:’SAR', type:’Indicator’, subType:’SAR’, definition:'The point estimate for smolt-to-adult return rate, calculated as 100 X the point estimate of the number of returning natural origin adults, divided by the point estimate of the number of smolts that produced those returning adults.', required:'Required if NullRecord = "No". Express these values as percentages (numbers from zero to one hundred), with two digits to the right of the decimal point. Examples: .020 = 2.00, .0015 = 0.15. This field holds a numeric value only -- the percent sign is implied but not included. Do NOT include repeat spawners in the number of adult returns. (A fish only returns once from smolting; subsequent returns are not appropriate for inclusion in smolt-to-adult estimates because they head to sea as adults on subsequent trips and thus are not exposed to the same suite of mortality factors.)’, sourceName:’Coordinated Assessments Data Exchange Standard Version 20170701’,sourceURL:'https://www.streamnet.org/coordinated-assessments-des/'})

The numeric value for SAR, with closely related numbers (SARUpperLimit, SARLowerLimit, etc.) will be packaged as a node. That node (the data) is linked to nodes for who, where, etc. (the metadata). It is also linked to ***Idea***s about SAR including definitions for methods, for taxonomy, for the name SAR (as above) and for all related names.

If we stick to the relational database structure for a SAR observation, all of these fields would be in a SAR record:

ID, CommonName, Run, RecoveryDomain, ESU_DPS, MajorPopGroup, PopID, CBFWApopName, CommonPopName, PopFit, PopFitNotes, SmoltLocation, SmoltDef, SmoltLocPTcode, AdultLocation, ReturnDef, SARtype, ScopeOfInference, OutmigrationYear, TRTmethod, ContactAgency, MethodNumber, BestValue, SAR, SARLowerLimit, SARUpperLimit, SARAlpha, ReturnsMissing, ReturnsMissingExplanation, RearingType, TSO, TSOLowerLimit, TSOUpperLimit, TSOAlpha, TAR, TARLowerLimit, TARUpperLimit, TARAlpha, HarvestAdj, OceanHarvest, MainstemHarvest, TribHarvest, BroodStockRemoved, ProtMethName, ProtMethURL, ProtMethDocumentation, MethodAdjustments, OtherDataSources, Comments, NullRecord, DataStatus, LastUpdated, IndicatorLocation, MetricLocation, MeasureLocation, ContactPersonFirst, ContactPersonLast, ContactPhone, ContactEmail, MetaComments.

Yuck. The beauty of re-creating these lugubrious standards as a graph is that many of these fields are replaced by links, so this SAR estimate is linked to specific 'Run', ‘SmoltLocation’,. And SAR is linked to its definition (idea and context for that idea). Less lugubrious (?), less redundant. And advantage is ease to to add or modify the kinds and content of the nodes, links, and ideas involved. For instance, we could have a node type ‘metadata’ so each node in a time-series of SAR estimates would be linked to one instance of that metadata. More than one metadata node would be required if methods changed.

Seriously. We should recast existing standards as graphs before we proceed. Current practices are paralyzing, suffocating. I would have greatly preferred to ignore all of this, but I had to wade in, thinking that standardization is absolutely required as a deliverable from IYS-SST in order to proceed with non-trivial integration (beyond Canadian).

The good news is that we don’t have to mobilize the standards for everything to get started. We can get started with just about nothing, then expand and revise as subsequently required. Including building extensions to whatever standards we start with. But we need to plan this with care.

Rationale / Goal

The datasets, and the data in knowledge graphs, assembled for ISDL examples, need to be understandable and useable by diverse users without explanations by the ISDL participants.

Project Objectives

ISDL graphs contain nodes and links that provide compliance with the Darwin Core protocols.

Background

Metadata

https://www.nceas.ucsb.edu/news/a-search-and-rescue-team-for-salmon-data

From the Darwin Core guide: http://rs.tdwg.org/dwc/terms/guides/rdf/index.htm :

- “Darwin Core is a vocabulary which provides terms that can be used to describe the properties and types of entities (known in RDF [resource description framework] as ‘resources’) in the biodiversity realm.
    - RDF is a network of relationships, in contrast to typical database tables. …
    - RDF is intended to facilitate data and metadata discovery by machines ("clients"). …

    - Resources must be identified using standardized and globally unique identifiers known as internationalized resource identifiers (IRIs).
    - RDF users adhere to … conventions about identifiers, data transfer protocols, and application of vocabularies.
- Darwin Core is a general purpose vocabulary because its terms can be used as part of a number of data transfer systems.
- It is assumed that **data from one provider will be linked to data from other providers**. This also implies that it is always possible to discover new data properties about a particular resource and that those properties may be described using unfamiliar terms. This differs significantly from other data transfer systems where there must be a pre-existing agreement between the sender and receiver …
- Anyone can make statements about a resource without agreeing to a pre-determined schema. ”

Hacking

Darwin Core http://rs.tdwg.org/dwc/terms/#dcindex Daunting. This is “**how**” to create an ontology, not “**what.**” We will need an actual, filled-in, extensive, usable ontology as the foundation for standardizing datasets so they can be integrated. I am looking for such a thing, perhaps NCEAS and SASAP. Else ... big job to fill in a lot of places, ideas, taxa (taxonomy),. But maybe that will be foundational data.

http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp

34,904 species (and more every year) and about that many synonyms, before getting to an ocean of common names in many languages.

https://www.calacademy.org/scientists/catalog-of-fishes-classification/ … begging to be a graph:

Class Actinopteri
Order Salmoniformes
Family Salmonidae Jarocki / Schinz 1822 (salmonids)
Subfamily Salmoninae Jarocki / Schinz 1822 (salmons, trouts, chars and allies)

This is getting closer to what we need:
http://www.fishbase.org/Nomenclature/ScientificNameSearchList.php?crit1_fieldname=SYNONYMS.SynGenus&crit1_fieldtype=CHAR&crit1_operator=EQUAL&crit1_value=Oncorhynchus&crit2_fieldname=SYNONYMS.SynSpecies&crit2_fieldtype=CHAR&crit2_operator=CONTAINS&crit2_value=&typesearch=simple&group=summary&backstep=-2&sortby=validname

I hacked out an excel sheet where the 16 scientific names for *Oncorhynchus* species are the same as the valid species name, following. How to separate the hyperlink from the text in a cell is a problem.
Does this mean building extensive ontologies as Google Sheets designed for easy conversion to a graph, including links? There are tools to convert a Google Sheet to Neo4j directly. Somebody has done this already, surely.

Oncorhynchus Species (15)

Scientific Name	Author	Valid Name	Family	English Name
Oncorhynchus aguabonita	(Jordan, 1892)	Oncorhynchus aguabonita	Salmonidae	Golden trout
Oncorhynchus	(Miller, 1972)	Oncorhynchus	Salmonidae	Apache trout

apache		apache		
Oncorhynchus chrysogaster	(Needham & Gard, 1964)	Oncorhynchus chrysogaster	Salmonidae	Mexican golden trout
Oncorhynchus clarkii	(Richardson, 1836)	Oncorhynchus clarkii	Salmonidae	Cutthroat trout
Oncorhynchus formosanus	(Jordan & Oshima, 1919)	Oncorhynchus formosanus	Salmonidae	
Oncorhynchus gilae	(Miller, 1950)	Oncorhynchus gilae	Salmonidae	Gila trout
Oncorhynchus gorbuscha	(Walbaum, 1792)	Oncorhynchus gorbuscha	Salmonidae	Pink salmon
Oncorhynchus iwame	Kimura & Nakamura, 1961	Oncorhynchus iwame	Salmonidae	Iwame trout
Oncorhynchus kawamurae	Jordan & McGregor, 1925	Oncorhynchus kawamurae	Salmonidae	
Oncorhynchus keta	(Walbaum, 1792)	Oncorhynchus keta	Salmonidae	Chum salmon
Oncorhynchus kisutch	(Walbaum, 1792)	Oncorhynchus kisutch	Salmonidae	Coho salmon
Oncorhynchus masou	(Brevoort, 1856)	Oncorhynchus masou	Salmonidae	Masu salmon
Oncorhynchus mykiss	(Walbaum, 1792)	Oncorhynchus mykiss	Salmonidae	Rainbow trout
Oncorhynchus nerka	(Walbaum, 1792)	Oncorhynchus nerka	Salmonidae	Sockeye salmon
Oncorhynchus rhodurus	Jordan & McGregor, 1925	Oncorhynchus rhodurus	Salmonidae	Japanese Amago
Oncorhynchus tshawytscha	(Walbaum, 1792)	Oncorhynchus tshawytscha	Salmonidae	Chinook salmon

A Network of Ideas Related to Salmon

- toward a standardized cross-index for salmon ecology.

A cross-index for books is a old idea, but one that is evolving. Consider the concept of "faceted classification" wherein you consider the most important dimensions of the ideas space for your subject, and ideally these dimensions are orthogonal (independent, do not share information). Facets, as in the reflections from a finished diamond that you are turning in your hand, refers in this case to looking at something from differing perspectives. What would those be for salmon ecology? Should we include everything— science, harvesting, habitats, people, and the management of all these? Sure, WTH (I confess to thinking about this since 2006). Try this:

Primary Facets

Ecosystem Structure - (taxonomy, geographic features, habitats, boundaries; *the pieces.*) consumed by DC: location, taxon, organism,

1. **Ecosystem Function** - growth, migration, predation, competition, development; *how things work*.
2. **Management and Finance** - plans, budgets; *how we work.*
3. **DataSet -** various formats, URL, metadata, (see #10 re one datum);
4. **Analysis and Knowledge** - knowledge management, models, workflows; *how we know.*
5. **Practice** - technology, methods, practices, analyses; *how we know.*

6. **Activities -** meetings, field work, operations, research, collaboration, (not event); *what we do.*
7. **Person -**
8. **Organization** - institution, agency, corporate/government structure, committees, teams;

To be compatible with Darwin Core, their eight facets are included. The field names are indicate content.

9. **RecordLevel-** institutionID, collectionID, datasetID, institutionCode, collectionCode, datasetName, ownerInstitutionCode, basisOfRecord, informationWithheld, dataGeneralizations, dynamicProperties")
10. **MeasurementOrFact** - "measurementID, measurementType, measurementValue, measurementAccuracy, measurementUnit, measurementDeterminedBy, measurementDeterminedDate, measurementMethod, measurementRemarks"
11. **Taxon** - "taxonID, scientificNameID, acceptedNameUsageID, parentNameUsageID, originalNameUsageID, nameAccordingToID, namePublishedInID, taxonConceptID, scientificName, acceptedNameUsage, parentNameUsage, originalNameUsage, nameAccordingTo, namePublishedIn, namePublishedInYear, higherClassification, kingdom, phylum, class, order, family, genus, subgenus, specificEpithet, infraspecificEpithet, taxonRank, verbatimTaxonRank, scientificNameAuthorship, vernacularName, nomenclaturalCode, taxonomicStatus, nomenclaturalStatus, taxonRemarks"
12. **Identification** - "identificationID, identificationQualifier, typeStatus, identifiedBy, dateIdentified, identificationReferences, identificationVerificationStatus, identificationRemarks"
13. **Location** - "LocationID, higherGeographyID, higherGeography, continent, waterBody, islandGroup, island, country, countryCode, stateProvince, county, municipality, locality, verbatimLocality, minimumElevationInMeters, maximumElevationInMeters, verbatimElevation, minimumDepthInMeters, maximumDepthInMeters, verbatimDepth, minimumDistanceAboveSurfaceInMeters, maximumDistanceAboveSurfaceInMeters, locationAccordingTo, locationRemarks, decimalLatitude, decimalLongitude, geodeticDatum, coordinateIUncertaintyInMeters, coordinatePrecision, pointRadiusSpatialFit, verbatimCoordinates, verbatimLatitude, verbatimLongitude, verbatimCoordinateSystem, verbatimSRS, footprintWKT, footprintSRS, footprintSpatialFit, georeferencedBy, georeferencedDate, georeferenceProtocol, georeferenceSources, georeferenceVerificationStatus, georeferenceRemarks"
14. **Event** - "eventID, parentEventID, fieldNumber, eventDate, startDayOrYear, endDayOrYear, year, month, day, verbatimEventDate, habitat, samplingProtocol, sampleSizeValue, sampleSizeUnit, samplingEffort, fieldNotes, eventRemarks"
15. **Organism** - "organismID, organismName, organismScope, assocatedOccurrences, associatedOrganisms, previousIdentifications, organismRemarks"
16. **Occurrence** -"occurrenceID, catalogNumber, recordNumber, recordedBy, individualCount, organismQuantity, organismQuantityType, sex, lifeStage, reproductiveCondition, behavior, establishmentMeans, occurrenceStatus, preparations, disposition, associatedMedia, associatedReferences, associatedSequences, assiatedTaxa, otherCatalogNumbers, occurrenceRemarks"

If we started with just the salmon world, we might have lumped Organism (apparently one observation), Taxon, and Identification (apparently method). It may be a mistake to smoosh together two separate communities of thought. Where do all of the IDs and standards come from? Which of many formats for date in Event? For instance, there are a vast number of categories and names for “lifeStage.” I see loose ends, such as catalogNumber without a catalogID.

Should we entirely split ideas from things? A thing is material, against which you can strike your foot with a mighty force. Person, Organization, Activity, and Place are things we want to discover through ideas and to interrelate by ideas. You cannot drop "taxonomy" on your foot, but how about "arctic" and "*Salmo salar*?" But if the objective is to standardize every name and every practice in order to build a single, all-encompassing knowledge graph for salmon, then

Is that enough? Better to split than be ambiguous. Should the following be separate dimensions, separate tree-trunks? Analysis, Communication, Organization,

Overall Structure

The are large issues to be addressed regarding "analysis ready data" including interpretable metadata, standardized ontologies (names), and indexing by a network of ideas standardization, and
The following paper was sufficiently interesting to hack their table of recommendations into this document.
I would go a step further, and ~~insist~~ recommend that the people who collected the original data receive "ongoing attribution" at all stages of the information flow including analysis results such as graphs, reports, and refereed papers. Which would require that all work be done in a "network aware" environment that (automatically) enabled copying the links from data to collector to be links from data product to collector. Scientific authors now have planet-wide-unique IDs. The most humble of data providers need to be recognizable and recognized in exactly the same way. Who will have made the more lasting contribution, 20 years from now: the author of a paper that is no longer read or cited (with ~99% probability) or the field biologist whose dataset contributes to a time series that is continuously updated and frequently re-analyzed?

Gil, Y., et al. (2016), Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance, *Earth and Space Science*, 3, 388–415, doi: 10.1002/2015EA000136.
https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2015EA000136

Table 1. A proposed checklist for gpf authors, with 20 recommendations that can guide them to assemble the information that should be included in a GPF [geoscience paper of the future].

Category	Applicability	Recommendations
Data accessibility	Initial data, significant intermediate results, and final results	D1: Data sets should be published in a publicly accessible location with a permanent unique identifier D2: Data sets should have a license D3: Data sets should be cited in the paper
Data documentation	Initial data, significant intermediate results, and final results	D4: Data sets should have general‐purpose metadata specified D5: Data set characteristics should be explained in detail D6: Data set origins and availability of related data sets should be documented
Software accessibility	Software used to process initial data and to generate any intermediate or final results	S1: Software should be published in a publicly accessible location with a permanent unique identifier S2: Software should have a license S3: Software should be cited in the paper
Software documentation	Software used to process initial data and to generate any intermediate or final results	S4: Software function and purpose should be described S5: Software download and execution requirements should be documented S6: Software testing and reuse with new data should be documented S7: Software support for extensions and updates should be mentioned

Provenance documentation	Provenance of all computational results reported in the article, including figures, tables, and other findings	P1: Derivations of newly generated data from initial data should be provided P2: Software execution traces for newly generated results should be provided P3: Versions and configurations of the software should be specified P4: Parameter values used to run the software should be specified
Methods documentation	Computational methods that are generally applicable to data other than the data in the paper	M1: Compositions of software that form a general reusable method should be specified M2: Data flow across software components should be described
Authors Identification	Authors of the paper and of any new data and software cited in the paper	A1: Authors have a permanent unique identifier

2. Spawner Surveys

Rationale / Goal

Project Objectives

Background

A link to the NuSeds database was posted: >390,000 annual reports of salmon spawner abundance (5 species) from > 4,000 spawning sites in British Columbia. Another database has details about the observations within annual estimates (but not for all). Matching subsets of both dataset will be converted and linked as a single graph.

Conventional analysis is "area under the curve" -- integration of total abundance from individual counts within a site within a year. Among many new analyses possible, we can describe the distribution of spawner abundance with time (the curve) in the context of many years and many sites. Determining the precision of each annual estimate would be a step forward. As a distant glow, knowing how the shape and location of the spawning curve changes due to river flow and ocean temperatures (among myriad habitat indicators) will allow predicting the total abundance with a year from the first few observations. Sounds like a refereed paper. Getting quality control feedback to the collectors during the collection (thereby quality assurance) is, perhaps, another project.

3. Web-based Mobile Tools

Rationale / Goal

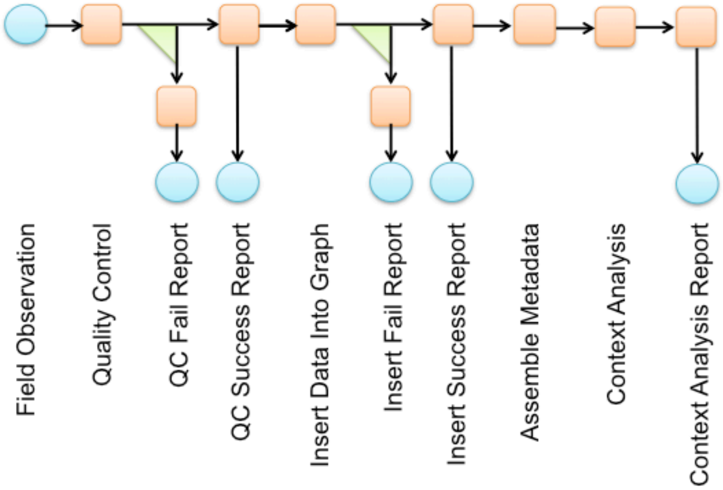
Project Objectives

Background

4. Workflow

Rationale / Goal
Non-technical staff need to use and adapt sophisticated tools without writing code. Thus automated data processing with decisions, a workflow.

Project Objectives
Example: upload a standardized dataset which is a day's observation in the field, say a spawner survey; subject that to QC and report, if OK the add the data to a larger assembly, say a graph; if OK then assemble related required for analyzing the new data in context; analyze, report back to the collector.



Background
Looks like some or all of this exists via GeoOptix from [Sitka Technologies](https://www.geooptix.com/home) in Portland OR.
https://www.geooptix.com/home

Decision trees and workflows **within** neo4j - https://maxdemarzi.com/tag/decision-tree/ this creates "path" nodes that execute via

5. Analytics

Rationale / Goal

Project Objectives

Background

6. Visualization

Rationale / Goal

Project Objectives

Background
Katy Börner http://info.ils.indiana.edu/~katy/ See her TEDx talk. Especially at 10:00 macroscope tools, describing information flow: load, clean, viz. https://twitter.com/hashtag/ivmooc
And her institute: https://cns.iu.edu// and newest book http://scimaps.org/atlas2
Atlas of Science. Atlas of Forecasting. Recommended by Bruce Hecht.
Examples of *Macroscopes for Making Sense of Science*: http://scimaps.org/iteration/12

The IVMOOC course outline looks like a roadmap for ISDL:

- Stakeholder needs acquisition & project specification
- Data mining algorithms and visualization tools
- Temporal, geospatial, topical, and network visualization techniques
- Research and development frontiers

And other bits such as:
R tutorial: how to identify communities of items in networks by Eiko Fried

7. Data Sharing Arrangements (Deals)

A next step for the IYS-SST work is actually sharing data, either by (a) making it public, (b) available to the workshop attendees (which will grow into a larger group of trusted affiliates), (d) to a team of co-authors, or (d) one-on-one. Making a deal about sharing data based on trust is one thing; recognizing it as a binding contract is another. Better safe than sorry. Example:
https://www.pnamp.org/document/coordinated-assessment-data-sharing-agreement-dsa-end-user-license-agreement-eula-and-data-use-policy-may-27-2016

Again, this can be simplified via graph database technology. All we need to do is define a secure workGroup. To join, you are vetted by the workGroup owner (also a member) and you accept the terms. Whereupon some invisible nodes become visible to you: datasets, papers in prep, proposals, budgets, other contracts,. Neo4j supplies the fine-grained security required: by node, by person, by file. We can investigate how to invoke and manage blockchain smart contracts so these deals are immutable.

7.1 Reusable Research

… table reduced by myself.

Table 1. Recommendations for authors to assemble the information that should be included in a geoscience paper. Derived from: *Toward the geoscience paper of the future: best practices for documenting and sharing research from data to software to provenance.* AGU 2016. DOI: 10.1002/2015EA000136 ResearchGate

Data	D1: Data sets are published in a publicly accessible location with a permanent unique identifier
	D2: Data sets have a license
	D3: Data sets are cited in the paper
	D4: Data sets have general-purpose metadata specified
	D5: Data set characteristics are explained in detail
	D6: Data set origins and availability of related data sets are documented
Software	S1: Software are published in a publicly accessible location with a 'permanent unique identifier
	S2: Software have a license
	S3: Software are cited in the paper
	S4: Software function and purpose are described
	S5: Software download and execution requirements are documented
	S6: Software testing and reuse with new data are documented
	S7: Software support for extensions and updates are mentioned
Provenance	P1: Derivations of newly generated data from initial data are provided
	P2: Software execution traces for newly generated results are provided
	P3: Versions and configurations of the software are specified
	P4: Parameter values used to run the software are specified
Methods	M1: Compositions of software that form a general reusable method are specified
	M2: Data flow across software components are described
Authors	A1: Authors have a permanent unique identifier

References

Börner, Katy. 2010. Atlas of Science. The MIT Press. ISBN:0262014459 9780262014458 https://dl.acm.org/citation.cfm?id=1995300

Börner, Katy. 2015. Atlas of Knowledge: Anyone Can Map.224 pp. MIT Press. 978-0-262-02881-3

Darwin Core. 2014. Darwin Core maintenance group, Biodiversity Information Standards (TDWG). https://doi.org/10.5281/zenodo.592792

Hilborn, R. 1985. Simplified Calculation of Optimum Spawning Stock Size from Ricker's Stock Recruitment Curve. Canadian Journal of Fisheries and Aquatic Sciences 42(11):1833-1834,. https://doi.org/10.1139/f85-230

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715 .

Glossary

HHGTTG - Hitchhiker's Guide to the Galaxy; literature
ISDL - International Salmon Data Laboratory; organization
IYS - International Year of the Salmon; project, inter-governmental organization
NASCO - North Atlatic Salmon Conservation Organization; inter-governmental organization
NCEAS - National Centre for Ecological Analysis and Synthesis; organization
NPAFC - North Pacific Anadromous Fish Commission; inter-governmental organization
NuSEDS - New Salmon Escapement Data System; database
OBIS - Ocean Biodiversity Information System; practice
PDD - Probability Density Distribution; statistics
RDF - resource description framework; subject -verb-predicate, triplets; information format
SASAP - Southern Alaska Salmon and People; project
SR - Stock Recruit; prediction of recruits to a fishery from abundance of spawners; model
SLA -
TDWG - The Darwin Working Group; organization

Notes

Eina Ooka - Doing Data Science Right in Excel-Pervasive Electric Utilities

Clutter created by years of Excel usage is inescapable in the utility industry as are all the associated bad consequences. A few years ago our team decided to utilize R to break out of this bad practice with the goal of building: Accurate, Explainable, Adaptable, Scalable, Accessible and Reproducible models. We built up our Modeling platform (R with GitHub), Data Platform (SQL Server) and Deployment Platform (R shiny Server) dedicated for analytics processes. With very limited help from IT developers, we -- a group of analysts -- had multiple obstacles to overcome along the way. In this talk I'll share our solutions to improving analytics workflow by adopting the latest R technology.https://earlconf.com/2018/seattle/#eina-ooka

---

Three Stages

There is a world of difference between the practical and pragmatic people who collect salmon data, whether field or lab, and the cerebral and abstruse people who want to illuminate ecological theories by using that data,
which means that communication between these two groups will be difficult (scant overlap),
BUT they both must have extraordinary faith in each other.

Field Biologist: This datum is worth my effort to collect because will be analyzed *forever* with increasingly valuable results (sort of interest on capital investment) by people who will respect my commitment and rigor.
Analyst: The data is worth analyzing because it was collected with rigor and consistency by careful, devoted, perhaps self-sacrificing biologists without whom this analysis -- which validates, honours, and repays their commitment -- would not be possible (and I would not be rewarded).
"The age of reason, based on faith." - Alfred North Whitehead

I do not propose Medieval Monastery as a model for IYS (attached). And yes, Father Gregr Mendel cheated when he recorded counts of sweet peas by colour. What I do propose is:

<u>Three stages and Three groups.</u>

- **Stage 1, Group 1.** Data identification and description. Who has what data? Where did that data come from, exactly? What do we need to know before we try to integrate and apply this data? To what other datasets is this data related (habitat indicators, biotraits, field notes, tags, methods,)?
- **Stage 2, Group 2.** Data assembly, cleaning, standardization, description (precision, reliability, categorization,). Database schema. Standardized glossary. Data transformation (renaming, reorganizing) and loading. Data integration. Preparation for analysis.
- **Stage 3, Group 3.** Analyses, visualization, new and more effect communication and decision support products. Refinement (feuds about) of hypotheses, models, and statistical practices. Scientific papers.

Then we do it all over again, with each cycle informed by, building on, and expanding the previous. Whilst, possibly, sporulating across the environmental sciences. Thereby "transformational."

Individual people can belong to more than one group, of course. The Group 1 people who will attend the 2019-01-23/24 workshop, are not exactly the original collectors, but represent them, and I suspect they want to participate in Group 3. At the very least they want to be sure of recognition (authorship, citation) by Group 3.

<u>What is different</u>
from previous conceptualization of the IYS workshops is:

#2 is new. We did not appreciate how difficult it will be to standardize the many different (sources X types) of data involved. This cannot be two points per year: stock, returns, *plug your ears, la-la-la-la-la.* But once you accept metadata, habitats, biotraits (fecundity, size), back-calculate length-at-age, then data integration is a large problem. Not to mention 21st century datasets: parental genetics that identify spawning groups (sub-populations within populations within CUs), genomic signatures (for disease, starvation, and stress),. I think this assembly and integration, however piece-meal our examples by 2022, is the way out of the box.

#3 lumps the previous. We split analyses from communication. After the collection of raw data, 95% of the work is to get the data to the point where it is analyzable. Running R scripts for statistical analyses, automated reports, or to create interactive visualizations is not what is holding back progress. Plus the analyses and communication part is enormously diverse, idiosyncratic, and the work of small teams – they will have escaped from the box. analyses and communication are ***enabled*** by the data standardization and integration. If you look inside the term “decision support products for decision makers” you discover everyone: fisheries and habitat managers, DFO and NMFS bureaucrats, politicians, ENPOs, First Nations and Tribes, local/state/provincial/federal governments, and everyone who votes. The analyses to create scientific results are a subset of the analysis to create decision-support products, especially upon realizing that scientists are decision-makers about hypotheses.

https://commons.wikimedia.org/wiki/File:Sumela_From_Across_Valley.JPG

(The Pacific Biological Station is an accretionary structure like this 4th century monastery, and has a separate wooden building – “The Annex” – where Jim Irvine, Kim Hyatt, and Scott Akenhead have their offices.)

Reply re three stages

From Jim Irvine: I like your minor epiphany with the three stages and can see how this could be presented effectively in a series of flow chart diagrams.

But first, to be honest, I am not sure how much faith the field bio has in the analyst. I suspect many are distrustful in how analysts might use their data. And the analyst typically has far too much confidence in the value of the data, in part because if they were not involved in data collection, they are unaware of all the warts and wrinkles. Also, as I believe Paul Smaldino said, analysts are often “lazy”, or at least unwilling to try to properly understand the data, or feel this is not their job. Lots of evidence for this.

The bridge between the Groups 1 and 3 is better communication and collaboration. [emphasis by Scott] This is what you are saying I believe when you suggest there needs to be overlap between people in Groups 1 and 3.

But Stage 1 Group 1 needs focus. Participants cannot simply identify/inventory data. There is too much data regardless of computing power. The issue is what types of data need to be identified, which is where hypotheses/mechanisms come in. In our case we are trying to predict the future by understanding the past. To do this, we need to know why things have changed. Do we need to measure temperature, cloud cover, and wind every time we go to the field? Probably not if it takes away from more valuable data collections. And even if we do collect this information, do we need assemble/present it? Not likely if it is not relevant to what we are trying to understand.

Stage 2 is the weak link where things fall apart. There is almost never enough time/support to this properly, which is ironic given the huge expense and effort typically allocated to Stage 1.

How can we elevate the profile of Stage 2? One simple idea is to present the flow of information from Stage 1 through a bottleneck at Stage 2 to get to Stage 3. In other words, show that the value of Stage 3 is directly dependant on Stage 2. The bigger the bottleneck (constriction), the smaller the ouput from Stage 3.

In reality, these are not totally distinct stages but rather overlapping processes that should have lots of feedback loops etc. I am already picturing a flow diagram where the degree of feedback decreased the constriction of the Stage 2 bottleneck thereby increasing Stage 3 output.

Revised Protocol for Conducting Recovery Potential Assessments

“in every case the best science advice possible with the information that can be assembled should be provided.“ complete Science support would include addressing at least the following questions:

1. What is the current status and trajectory of the species (or population)?
2. What are biologically reasonable recovery targets and timeframes to reach recovery for the species?
3. What features characterize the habitat of the species?
4. Where is the habitat found at present, how much habitat is known to exist currently, and how much habitat was known to exist historically?
5. What are the current threats to the species and its habitats?

6. What is the likelihood of reaching the biological recovery targets with current productivity and mortality rates estimated for the species?
7. What mortality rates and/productivities would be associated with alternative ways of conducting activities that affect the species?
8. By how much would various mitigation measures be expected to alter the mortality rate and/or productivity of the species?
9. Effectiveness of current management measures, if any are in place.

Phase I: Assess current/recent species status To the extent possible with the information available and taking account of uncertainties:
1. Evaluate present species status for abundance, range and number of populations.
2. Evaluate recent species trajectory for abundance, range, and number of populations.
3. Estimate, to the extent that information allows, the current or recent life history parameters for the species (total mortality [Z], natural mortality[m], fecundity, maturity, recruitment, etc.) or reasonable surrogates, and associated uncertainties for all parameters.
4. Address the separate terms of reference for describing and quantifying (to the extent possible) the habitat requirements and habitat use patterns of the species.
5. Estimate expected population and distribution targets for recovery, according to DFO guidelines.
6. Project expected population trajectories over three generations (or other biologically reasonable time), and trajectories over time to the recovery target (if possible to achieve), given current population dynamics parameters and associated uncertainties using DFO guidelines on long-term projections.
7. Evaluate residence requirements for the species, if any.

Phase II: Scope for management to facilitate recovery. To the extent possible with the information available and taking account of uncertainties:
8. Assess the probability that the recovery targets can be achieved under current rates of population dynamics parameters, and how that probability would vary with different mortality (especially lower) and productivity (especially higher) parameters.
9. Quantify to the extent possible the magnitude of each major potential source of mortality identified in the pre-COSEWIC RAP and considering information in COSEWIC Status Report, from DFO sectors, and other sources.
10. Quantify to the extent possible the likelihood that the current quantity and quality of habitat is sufficient to allow population increase, and would be sufficient to support a population that has reached its recovery targets (using the same methods as in step 4)
11. Assess to the extent possible the magnitude by which current threats to habitats have reduced habitat quantity and quality.
Phase III: Scenarios for mitigation and alternative to activities To the extent possible with the information available and taking account of uncertainties:
12. Using input from all DFO sectors and other sources as appropriate, develop an inventory of all feasible measures to minimize/mitigate the impacts of activities that are threats to the species and its habitat (steps 9 and 11).
13. Using input from all DFO sectors and other sources as appropriate, develop an inventory of all reasonable alternatives to the activities that are threats to the species and its habitat (steps 9 and 11), but with potential for less impact. (e.g. changing gear in fisheries causing bycatch mortality, relocation of activities harming habitat).

14. Using input from all DFO sectors and other sources as appropriate, develop an inventory of all reasonable and feasible activities that could increase the productivity or survivorship parameters (steps 3 and 8).
15. Estimate, to the extent possible, the reduction in mortality rate expected by each of the mitigation measures in step 12 or alternatives in step 13 and the increase in productivity or survivorship associated with each measure in step14.
16. Project expected population trajectory (and uncertainties) over three generations (or other biologically reasonable time), and to the time of reaching recovery targets when recovery is feasible; given mortality rates and productivities from 15 that are associated with specific scenarios identified for exploration. Include scenarios which provide as high a probability of survivorship and recovery as possible for biologically realistic parameter values.
17. Recommend parameter values for population productivity and starting mortality rates, and where necessary, specialized features of population models that would be required to allow exploration of additional scenarios as part of the assessment of economic, social, and cultural impacts of listing the species.

Dear Dr. Radchenko,

Perhaps you are aware of an initiative to design next generation practices for salmon data processing: the International Salmon Data Laboratory (ISDL), associated with the International Year of the Salmon (IYS).
I have organized a workshop 2019-01-25 to organize and advance the ISDL, please see the attached announcement (draft).

Because this is a new initiative, and despite intending "international," I am worried that participation from outside of Canada is likely to be scant.
If we are serious about integrating salmon datasets from around the world, and working together on the analyses made possible by such integrated data, then we need to understand the issues that will arise as we begin sharing data between institutes and countries. NPAFC has successfully delivered great value by catalyzing the sharing of summary statistics, and I believe there are a series of success stories about NPAFC catalyzing the sharing of detailed data: tag returns, otolith marks, age and length distributions,.

The intention of ISDL is the assembly, integration, and analysis of the data that lies behind the summary statistics. This detailed data is required to determine relative precision (from sample sizes, methods) of data points, investigate effects from climate and habitat changes, and to understand the threats and vulnerabilities to salmon at all life stages. The overall goal of the ISDL is that of the IYS: a mechanistic understanding of the reactions of salmon populations to global warming, and, based on that understanding, improve the practices of habitat and fisheries managers to enhance the resilience of salmon. IYS has brilliantly recognized that resilience for salmon means nimble and fully informed decisions, reacting to surprises with all of the knowledge we can muster. ISDL takes the next logical step, recognizing that attaining this goal requires a jump in the efficiency of data processing, analysis, and communication. While I lack confidence about delivering wisdom to decision-makers, we are certainly able to design computer systems that will deliver up-to-date information and knowledge.

Massive integration of salmon data is perhaps a goal for the generation that will replace us, but we can immediately design and experiment with computer systems that will allow such a thing. We are well into the 21st century:

technology limitations are not what is holding us back. *Immediate goals* for ISDL involve creating examples of what can be done with modern technology; creating and introducing new data-working tools for ecologists; and helping everyone see rewards from new practices for data processing. *Specifically,* the ISDL will support the work and the goals expressed by the preceding and subsequent IYS Salmon Status and Trends workshops. Focus —> Excellence —> Adoption.

Is there a possibility that you could attend the ISDL workshop? We badly need your international perspective, and your familiarity (I am being assumptive here) with data management practices and policies at TINRO.
If you could attend, that would be wonderful. You will see that I have gone beyond assumptive to be presumptive and suggest a topic that describes how you might contribute. You may wish to fix that. The topic, I mean.

Perhaps there will be representatives from TINRO in Vancouver for IYS meetings earlier that week. If so, perhaps some of them might be interested in this workshop.
Please feel free to forward this announcement to whomever you think might be interested in the ISDL.

Yours respectfully,

To AMH re COSEWIC funds

What I have in mind is setting up "streams" of data from original sources, that "flow" into a data "lake" where they are integrated and ready for analysis. 95% of the work is preparation for analysis: assembly*, cleaning, standardization, and integration.   *Analysis: less slog, more fun, all the credit.*

If we do this right, delivering the work PLUS building and ***applying*** the prototypes will be less work than delivering the work MINUS new efficiencies.

In my vision, the data lake is a knowledge graph (in neo4j with related libraries and tools) that links observations of abundance by life-stage to: metadata, bio-traits, habitat indicators,. Ideally, the **provenance** of summarized/analyzed data is maintained: where exactly did that *estimate* of mean length, or that *estimate* of spawner abundance come from? To what extent is that estimate precise, calibrated, and reliable? The good news is that it is easier (simpler) to extract a "pattern" of extensively cross-linked data from a knowledge graph than from a relational database of similar complexity (cypher queries within R scripts and within Javascript GUIs).

Ecology is complicated. Fine, embrace that complexity, that's our job. "Stock assessment biologists" and "fisheries harvest managers" need to see themselves as ecologists: knowledgable, responsible, skeptical, and scientific.

Pacific State of the Salmon Program

Fisheries & Oceans Canada's (DFO) new Pacific State of the Salmon Program was initiated in March 2017. The goal of this program is to develop tools and processes to foster salmon-ecosystem integration. Through this integrative work, Canadian Pacific salmon population and freshwater & marine ecosystem trends will be tracked and compared to better understand salmon status and the factors contributing to their statuses. Key salmon characteristics that will be included are abundance, productivity, body size, fecundity, and status.

The State of the Salmon Program relies on three key pillars to achieve these goals: **an interactive analytical tool, communication, and collaboration.** The analytical tool is foundational to this Program. [R and Shiny] This tool will enable users to select, match, compare salmon trends across populations, find overarching patterns and relationships. Approaches to synoptic overviews of Pacific salmon populations will also be developed. These outcomes will enable scientists and managers to answer key questions that support their research, monitoring, and management activities. The analytical tool will be foundational to both the communication and collaboration pillars of this Program. It will be used to communicate out broadly on salmon patterns and their relationships to one another, their ecosystems, and other factors contributing to their trends.

Other communication deliverables of this emerging Program include pre-season, in-season, and post-season reporting on salmon returns, escapements, and survival. The analytical tool will also be used to foster collaboration with experts on salmon and their ecosystems. There is considerable knowledge on salmon and their ecosystems among DFO staff across sectors, First Nations, and external groups that can be integrated within the analytical tool's framework. Other collaborative work will include comparing Canadian Pacific salmon population trends and status with salmon populations throughout the north Pacific, the Atlantic, and salmonid populations in the Arctic. This Program will facilitate an annual State of Salmon forum(s) to foster collaboration among experts on salmon and their ecosystems, and also deliver presentations and publications in a variety of forums.

Sue Grant, State of Salmon Program

Sue Grant is leading the new State of the Salmon Program in Fisheries and Oceans Canada's (DFO) Ecosystem Science Division. Her aquatic biology career in Canada over the past two decades has spanned Atlantic cod on the east coast, charismatic mini-fauna (fathead minnow) populations in Northern Alberta, and killer whales and Pacific salmon on the west coast. Sue has worked for DFO's High Seas Salmon Program, was the biological lead for catch monitoring assessments of among Canada's largest freshwater recreational salmon fisheries located on the Fraser River, and led stock assessments of Fraser River chum, sockeye and pink salmon. In the last decade, Sue's work focused on understanding population dynamics and the biological status of Fraser sockeye and pink stocks, which has emphasized integration of knowledge among experts on salmon freshwater and marine life-history and ecosystems. Sue will take all her past knowledge and expertise in salmon, ecosystems, analyses and communication and apply this to the implementation of the new State of the Salmon Program.

To PBS DB managers

We need to walk outside of our comfort zone into the data processing technologies of 2019… yes, you've lived into the future. A couple of concepts to whet your appetite:

1. modern data analysis works with all of the data at once. E.g. not spawner abundance estimated for one stream in one year (~5 complicated observations) but extracting the patterns of spawner abundance by day of the year, by stream, and by year; and then adjusting that well-defined pattern for each stream each year. This opens the door to using stream flows, local and regional sea temperatures,, to adjust the patterns. Hundreds of data points to estimate dozens of parameters, better than 5 data points for linear interpolation. Use everything we know to learn more. We have Etienne Rivot attending (I hope) who wrote https://www.amazon.com/Introduction-Hierarchical-Ecological-Environmental-Statistics/dp/1584889195
2. Linked to this is data integration. How do we merge multiple datasets for a suite of analyses? I think this is via graph database (think: mind map) technology, particularly neo4j. The main reasons are (a) there are tools to lift data into a graph and join graphs, (b) easy to add new types of information and new links between kinds of data without having to redesign the database, (c) simpler queries of a knowledge graph via Cypher than of a

complicated relational database via SQL, and (d) readily accessible via Rneo4j and similar for analysis and particularly for modern data visualization: interactive charts, infographics, dashboards,.

3. BUT standardization and an initial graph schema (basic types of nodes and links) lies in the way of data integration. Fine, something to get through.
4. Finally, we need to be more effective about data products for the diverse range of decision makers that impact (in this case) salmon fisheries and salmon habitat management. So, decision-support products. The lofty goal is nimble decision making, fully informed by everything we know based on up-to-date information.
5. All of which has been moved into crisis mode by the long-predicted and long-procrastinated arrival of anthropic global warming. It may be that the 20th century data is all irrelevant, this is a new, surprising, and dangerous world. Salmon have to be resilient, so management has to be nimble, so information flow has to be much better.
6. If we can’t save the salmon, how the hell are we going to save the humans?