Hello, I am Briana Sullivan. I work in the Data Visualization Lab on the "Chart of the Future" related projects.

I'd like to discuss the Coast Pilot data structure as it is today and some insights on where I think it should be and some ideas on how to make it happen.

I'll follow up with a brief synopsis of the work we are involved in with the various IHO Working Groups.

Currently

Based on the publication

Joint Hydrographic Center Annual Review • July 14–16, 2015

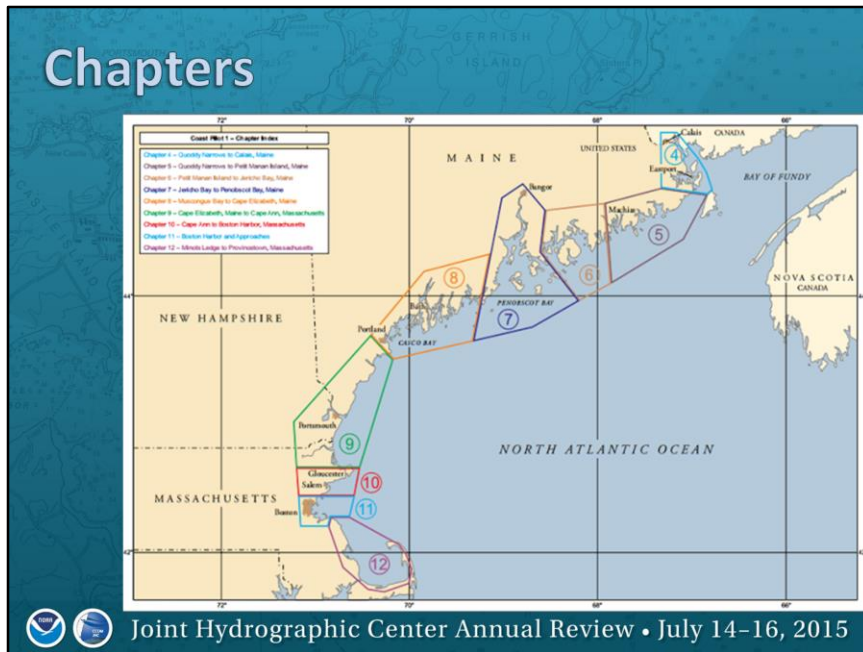let me first explain, visually, how the data structure is based on the publication…

Joint Hydrographic Center Annual Review • July 14–16, 2015

The Coast Pilot is set up via Books. Books subdivide the working area into maintainable sections. So, to get local information you need to know the Book as the root element

The chapters unique to each book, subdivide the area represented by the book into manageable parts. NOTE: each chapter contains one or more nautical chart within it's limits (and due to this some information repeats where charts/chapters/books share a border/overlap)

Within each book are various chapters, some with data common to the entire book, most chapters are for specific areas delineated by a specific boundary.

Clearly there are some chapters that don't follow the same type of formatting...these will be researched and dealt with at a later time.

Notice this disclaimer on the left that this is "...information that is difficult to display on a nautical chart." ...I'll talk about this a bit later.

Within each chapter are "elements". Each element is represented as a paragraph within the chapter and is numbered. But the elements have types associated with them such as:

Chapter Header (Orange)

Chart Header (Yellow)

Paragraph Header (Aqua)

Paragraph -> CP_Tags (Blue)

Image (Pink)

These items are typically described in geographic order. On the East Coast of the US, for example, it is from North to South.
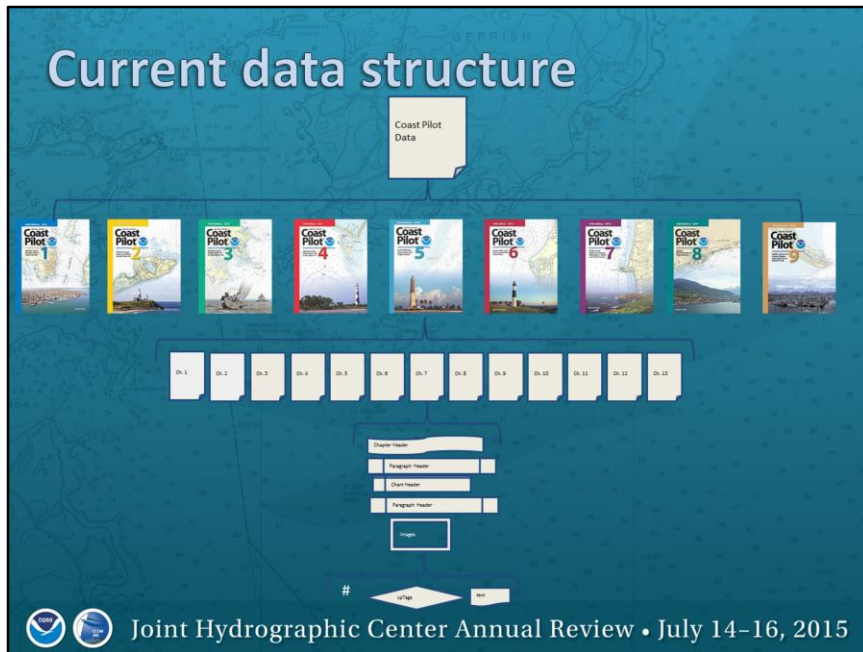
This is a hierarchal data structure ….just like a tree, where the whole CP is the tree, each book is a branch, each chapter are leaves on the branch, etc.

This is what the "tree" looks like when represented with XML. You can see the tags are exactly as you'd expect the layout of the book!

With the S-100 standard in the works, technology use common place, and a need for more useable and accessible data it's "time to think differently".

It's probably better not to think of Google Maps (or Nautical Charts) as a thing like a paper map.
Geographic information systems represent a jump from paper maps like the abacus to the computer.

A quote from London cartographic historian ...

(read quote at top of slide).

A former NASA engineer working at Google lays out the steps to the success of Google Maps...(read quote at bottom of slide)

Isn't that the holy grail? Getting something out of this that is "higher quality than the sum of its parts"?

source: http://www.theatlantic.com/technology/archive/2012/09/how-google-

builds-its-maps-and-what-it-means-for-the-future-of-everything/261913/

We can accomplish this if we can move from a publication-centric data format to a data-centric one. What do I mean by this?

Rewrite

"(Coast Pilots) contain supplemental information that is *difficult to portray* on a nautical chart."

Coast Pilots *describe* the features
found along the coast
and
how to get to / access / use / avoid them.

Joint Hydrographic Center Annual Review • July 14–16, 2015

It brings me back to that disclaimer I mentioned earlier…(read slide).

OCS is starting the process of thinking differently

From the OCS website

The Coast Pilot project needs to **share its information** with other sources and offer a *customizable* product.

This degree of customization requires *more control* than a traditional Desktop Publisher (DTP) can provide

…the Extensible Markup Language (XML) and associated tools have that capability.

If done correctly!

Joint Hydrographic Center Annual Review • July 14–16, 2015

…I recently found this on their web page…(read slide)

But remember…just tagging something with XML doesn't mean it will allow the kind of control mentioned here.

Thought and foresight needs to come into play to make sure the mark-up is being used to maximize data use.

source: http://www.nauticalcharts.noaa.gov/nsd/RD-cpdb.html

So to demonstrate how "Thinking Differently" could work to maximize the data…here's an example paragraph.

Notice how the info in the red outlines could be considered "metadata" to this chaper/book. The Chart headings would be irrelevant in a world of charts based on features. Now the line is blurred between charts and boundaries that were so prevalent in the past.

The XML colored in black is directly from paragraph 5 (the green highlight). Basically, I read the paragraph and attempted to take everything I could and "tag" and organize it. But, since I'm focused on the feature, I'm stopping at the first one I see, "St. Croix River". Then I search the rest of the chapter for "St. Croix River". My goal is to find and group all info about the "St. Croix River" that I can. The orange text is just that…info from the rest of the chapter about the St. Croix River. Wow, how great is it to be able to find out all about this river in one location!

Doing this exercise brought up quite a few other questions in quality control with consistency, formatting, and organization. Like:
1) Why isn't the first instance of "St. Croix River" (5) in bold? Instead it is bold in

(104).

2) Principal entrance is listed in (5) as "Head Harbour Passage", yet elsewhere in the chapter "Friar Roads" is also a principal entrance
3) what is the determining factor to having the geo-referenced links?
4) What aren't all the occurances within the text (i.e. Whitlocks Mill) all geo-referenced? I should be able to click on any one of those feature names and get the same result.

the current XML (publication-centric) mark-up...comparing this time against the data-centric mark-up example on the right.

Note: the St. Croix River (the green box) is inside the <paraText> tag hidden amongst a narrative string of words. Not easy to quickly find or extract.

Granted...it's a LOT of work! But the uses for how the data is marked up on the right is many times more useful than the original.

For those of you who don't like to look at code, another way to view and understand this would be like filling out a form online…
The example on the left allows for a narrative - a rich textual description. Easy for input purposes, but just not easy to process, discover and use on the output.

The example on the right, sure it's more complicated to set up, it takes longer to fill out...BUT, it forces the input to be formatted, tagged, organized, and verified on submit, which means there is no more/very little work to do on the output...

This is a great example of how to standardize the data content.
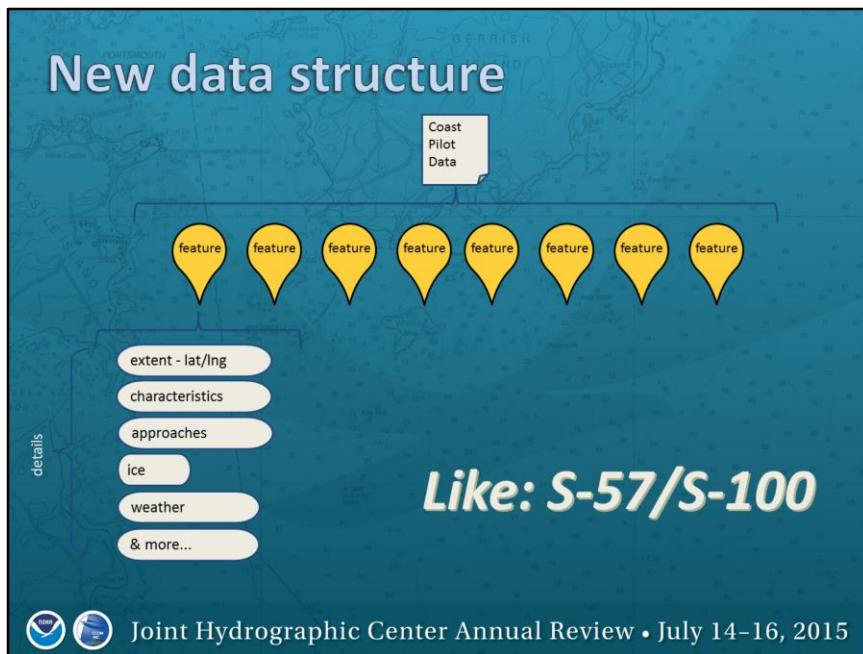
Yet another way to view this data structure (for those of you that like pictures!)

This is the direction of the S-100…Feature-based data…

it can work for the Coast Pilot/Sailing Directions too…in fact it should work in conjunction with S-101 (the ENC) data.

Everything related to a specific physical feature (S-126 - the physical environment) should be able to reference the S-101 ENC utilizing the overlap and simply linking the features together. Remember the goal to reduce redundant work and work towards "harmonization" with other related specifications.

For the most part the S-126 is basically describing the S-101 feature….the textual description of the physical feature as it relates to navigation.

Why? – the end result

1. A mariner of a small private vessel wants to get rid of all text relating to pilots and large ships. (and vice-versa)
2. Correlating the weather data in the text with historical, current and predictions of the same area.
3. A researcher wants to test AUV in a specific area and wants a quick way to search for all rivers that have a "windy" characteristic.
4. Filter out anchorages along a coastline to help a mariner decide which to head for. (using S101 feature catalogue rules of anchorage types relevant to her - code 8: small craft mooring area).
5. Get drawbridge information for a route quickly...all in one place.
6. Text reader for a audio description of area

Joint Hydrographic Center Annual Review • July 14–16, 2015

Remember, we need to keep in mind the end result, how will the data be used? It is (or should) be there to serve us...not us be slave to it.

1. (or better would be "winding" since "windy" in the dictionary is "wind swept" not "twisty"...but that's another presentation!)
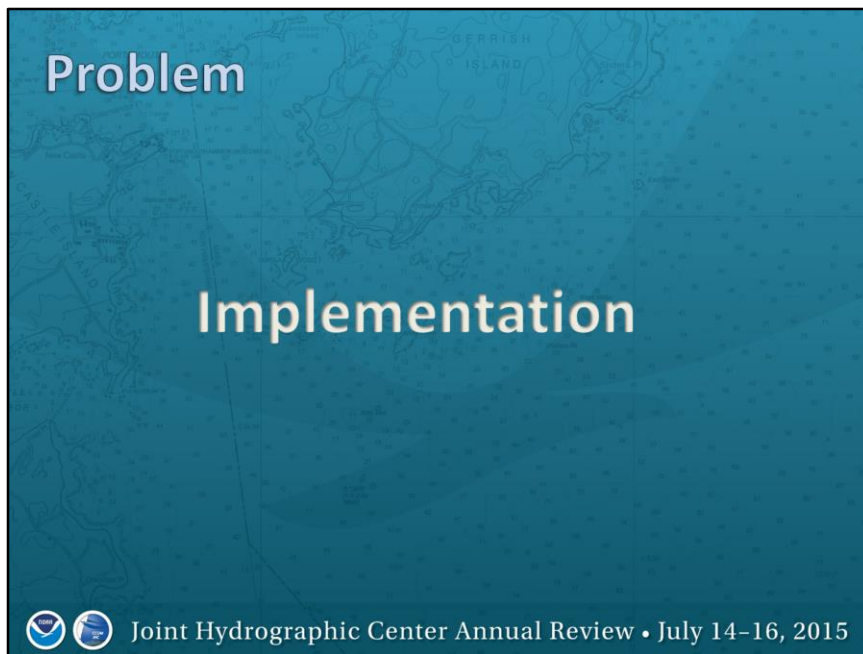
Of course, the biggest problem in this whole story is HOW do you get from the current situation to the dream?

Or even something on the way to the dream? I don't think the presentation should end here (even though you may want it to!)...

….So, how to make it happen? How do you turn a tree into lots of little seeds (for lack of better analogy)?

And the million dollar question how do you take the seeds and get the tree you are used to?

Intermediate steps
1. *Normalize*/clean the database
2. Geo-reference data (use GML)
3. Add Chart numbers to all elements (for backward compatibility)
4. Better Tags:
   1. Replace cp_index tags to ref/extend feature tags - *standardize*
   2. Replace "formatting" tags (cp_bold, cp_italic) with feature tags
5. Add categories for filterable data
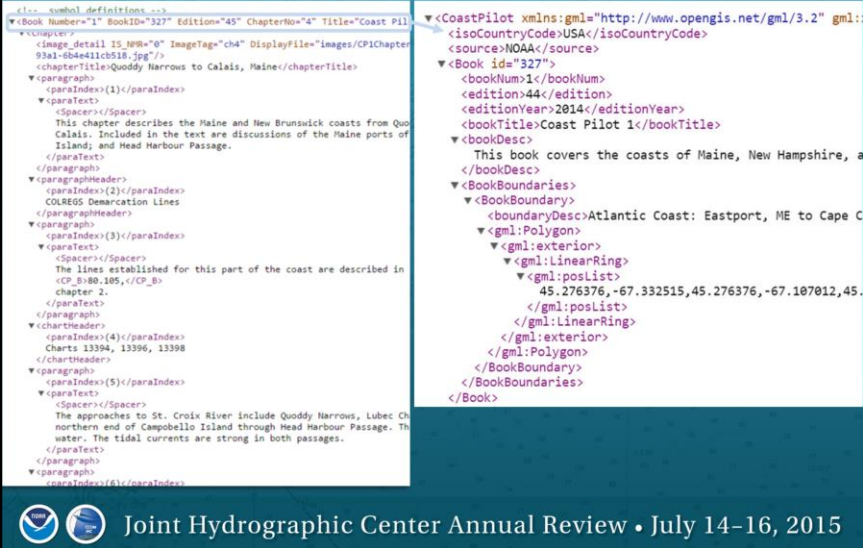6. Recode tags with fewer attributes

Joint Hydrographic Center Annual Review • July 14–16, 2015

1. Process of reducing redundancy in the DB.
   a. Use query string to find distinct Paragraph Header type of elements, then clean the list manually
      i. Get rid of tags, duplicates, make things generic (get rid of state and regional specific locations in text) - this could be a starter list for common headers in all of the world's sailing directions -- This is part of the STANDARDIZATION process! (example: Pilotage for Coast of New Hampshire, just Pilotage since it's already in the section about New Hampshire)
         1. duplicates in the database reduce headers by 318 entries (from 2258 to 1940),
      ii. This is a great exercise to see how consistent the document is written, how terms/phrases/headers are similar but not the same.(i.e. using MA vs. Mass, VA vs. Va., NY vs. N.Y., under way vs. underway, use of caps, use of spaces/punctuation, plural vs. singular, cross current vs. crosscurrent, small craft vs. small-craft vs. small-boat)
   b. Get rid of added tags and attach as a new field "type" or "paragraph header type" (Appendix, CFR, sect, CP_INDEX, CP_BOLD, etc.)

      c.   Punctuation within tags isn't necessary.
1. Geo-reference data - how is Tom's group doing it (regexp automation to start) According to publication "Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia"
    a. Checking for potential matches: find all possible matches (either use the xml version of each document or make new version from DB) - "feature catalogue"
    b. match with GNIS table
2. Isolate the the chart numbers using regexp, set up parent/child relationships via query and create new table of chart numbers to elements. This set will give everything an associated area even if it doesn't have a geo-reference.
3. Compare against the ENC feature catalogue so the tagging can be synced up/reused
4. Such as Anchorage, Pilotage, etc. Later these can be changed into feature categories that will comply with S-100 standards.
5. recode without attributes

Example: fewer attributes

Fewer Attributes: This is an easy interim step that could be done now to help improve the way the Coast Pilot XML is currently being done.
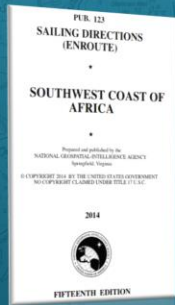
The data structure of the Coast Pilot

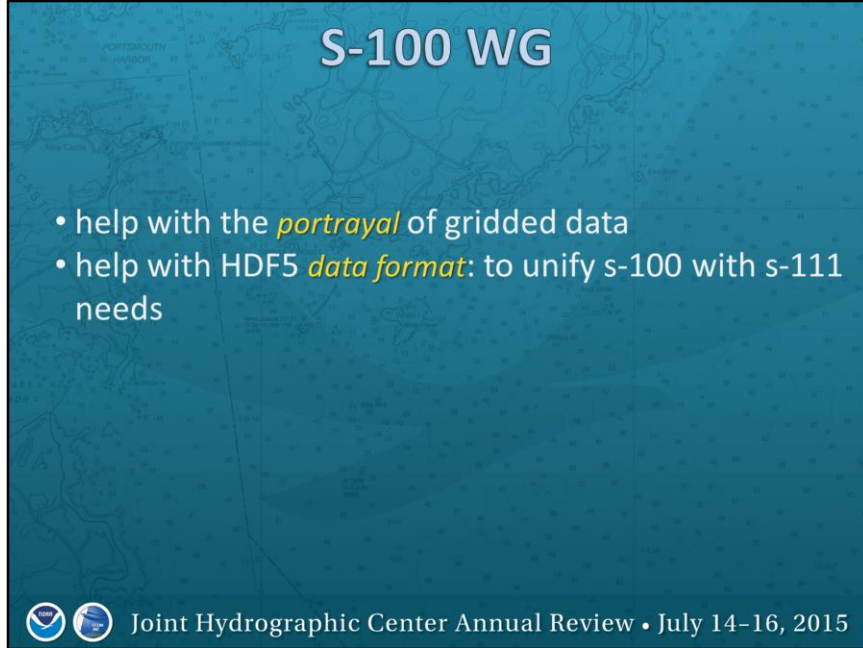Tides, Water level, Currents WG (TWCWG)
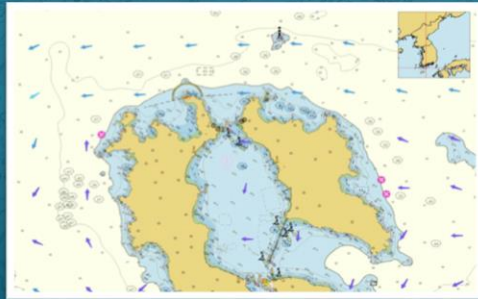
- 2014-2015 Develop initial draft of S-111 PS that includes *portrayal*, schema, and encoding

- 2014-2015 Test the *portrayal* of surface current vectors and real time current data

Joint Hydrographic Center Annual Review • July 14–16, 2015

The data structure of the Coast Pilot

The data structure of the Coast Pilot

Questions?

Joint Hydrographic Center Annual Review • July 14–16, 2015