NIPWG 6-05.3

Paper for Consideration by NIPWG
NOAA's NPB progress on transitioning from paper products to S-100 products
(progress on interoperability between CP and ENC)

Submitted by:	UNH and NOAA Office of Coast Survey – Nautical Publications Branch
Executive Summary:	A brief summary of the results of a workshop held to discuss interoperability and data structures within the NIPWG bailiwick.
Related Documents:	NIPWG1-06.3 (15 June 2015)
Related Projects:	S-126, MRN, interoperability with S-101

Introduction
Since 2015 when the IHO's SNPWG (Standardization of Nautical Publications Working Group) restructured and was renamed as the NIPWG (Nautical Information Provisions Working Group) it had been discussed amongst the University of New Hampshire (UNH) and NOAA's Nautical Publications Branch (NPB) at the Office of Coast Survey (OCS) that the Coast Pilot publication needed to be more data-centric if it is to keep up with the advances of the electronic chart. Making a comparison to the evolution of soundings on the chart, the publication-centric format of maritime documents are currently like "lead-lines" where a data-centric paradigm can transform them into "multi-beam". This paper is a report from a workshop led by UNH and held at NPB in November 2018 to begin the process of implementation of a data-centric production system; it will attempt to highlight the current publication-centric production system, the mechanism moving forward to slowly transition into a data-centric system, as well as discussions and issues that came up relating to interoperability with the ENC data. This paper is an attempt to share ideas and challenges with the global community on transitioning to a new data services system based on S-100 standards as well as to document the journey for others to learn from or lend ideas to.

Background
The US Coast Pilot's (CP) current production system involves a database laid out in a structure of tables containing the information and metadata for books, chapters and chapter elements (consisting of headers, paragraphs and images – see NIPWG1-06.3 for details). This database is then connected to an InDesign file that is able to read the contents of the database and populate the file to create each of the 9 books (9 geographical areas dividing up the coastal waters of the United States). This can then be printed out as a pdf, XML or HTML.

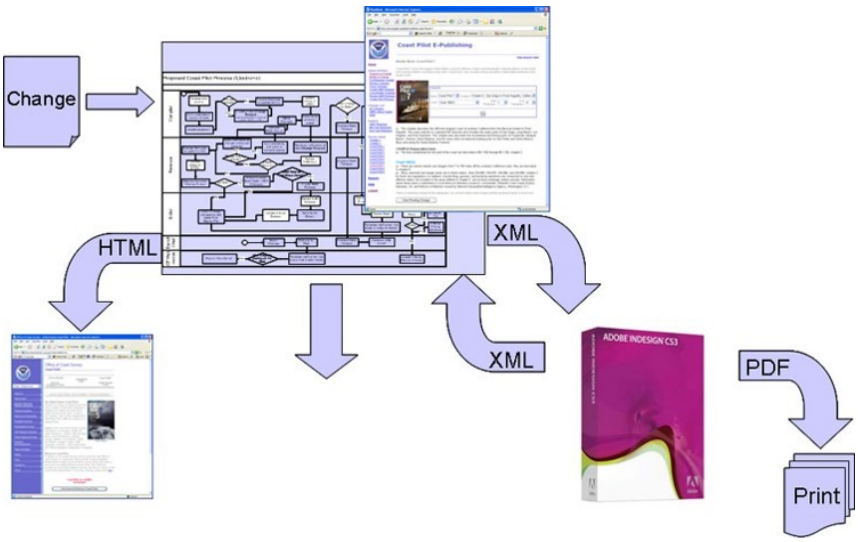

Figure 1. The current CP production system

The workshop was intended to work with this current production system (fig 1) helping to devise steps towards the evolution of a data-centric system that would ultimately produce and interoperate with S-100 products.

UNH brought to the workshop 2 years of trial and error testing with a static copy of the CP database (CPDB) and a recent acquisition of the US ENC database system. The experience UNH brought in connecting multiple datasets/databases together laid the foundation for a broad vision of the possibilities for a new system as well as a detailed discussion on issues discovered and a general plan for how to move forward. Building upon NPB's geo-tagging efforts UNH included mechanisms to add topics/categories to the text and also began structuring the data to prepare for generating IHO S-100 data products as well as see how it can add value to existing S-100 products.

Analysis/Discussion

This section summarises the main topics that were discussed over the course of the three days the workshop was held.

Geo-tagging Features

Within in the last 10 years or so the CP has made huge strides in geo-tagging the main features of the CP text with XML representing the point latitude/longitude values. UNH was able to extend this effort by automatically tagging *all named*/geo-referenced features (not just the indexed features – depicted in the CP with bold text). A new table was also created in the UNH test database that established relationships between all of the main features and all of the paragraphs where those features were found and tagged; this allowed for collating all information about one feature together, making the textual data more flexible in its presentation.

It was agreed upon at the meeting that a table would be created in the CPDB that would be populated with all the indexed geo-referenced features that had been tagged and become the first step in the paradigm change for the system. This would lay the foundation for CP features to be associated with the same features depicted on the ENC chart.

An issue came up for ambiguities in names (common names for features could be duplicated in other areas of the country) that would be resolved by checking against a Geographic Names Information System (GNIS) database that already had geo-referencing as well as other lat/lng points that are either contained within the same paragraph or surrounding paragraphs. Any not found in this automated way would need to be added by hand.

Topics/Categories

UNH also recommended to NPB to add a table that followed the Chart One categories and another with the associated subcategories. For example, Cultural Features would contain bridges and built-up areas, where Hydrography would contain Tides and Currents. The purpose behind this is to create a context for the CP data and prepare the way for separating out the data relevant to the S-100 product specifications (such as the Cultural Features will typically match up with S-101 features and the Tides with S-104 and Currents with S-111).
Use
The NPB started the process to normalize the headers in the CPDB and came up with a plan to reduce the redundancies in their system by adding another table to just list the common headers then reference those headers with id values to the header table instead of redundant text.

Since the Chart One categories is a well-established ontology already, leveraging that structure for organization accomplishes two tasks at once for NPB; since they were tasked a few years ago with taking over the maintenance of the Chart One publication.

Building Data Structures

The main exercise that was intended to be practiced at this workshop was the concept of reverse engineering the text within the CP into various data structures. This task involves an iterative process that would be accomplished by the subject matter experts who work with the CP books to begin to understand the data the CP had and how it could be better represented for machine use. UNH spent time in an iterative manner marking up Chapter 9 of book 3 as a test case. This area was chosen since it is the entrance to the Chesapeake Bay and coincides with data being generated for the S-111 (Surface Currents).

There are *many* different data structures within the CP data including anchorages, bridges, rules and regulations, various types of areas, and features including landmarks, as well as approaches, services and climatological weather and general water conditions. It was discussed that they should be worked on in the order of most importance and usefulness. One key way to determine this would be to run a statistical analysis on the CPDB based on the headers in each chapter (since the headers identify the topic for the context of the contents within it.) Determining which topics had the most changes over the lifetime of the database would help to see what is getting the most attention. It was also decided that some things might not have changed in decades (or ever) and to identify which things might not need the same attention or might not be needed at all. This would help to reduce the workload as well as identify out of date or non-value-added information.

(When we use the term “value-added” we mean to indicate adding support to what the chart already displays, which could include redundant data if it helps to assign a higher importance to specific ENC features). The goals for making sure the information that does become modernized for machine use would fulfil the following three main requirements:

- Enhances the user’s understanding of chart elements
- Broadens the user’s situational awareness
- Draws attention to the most important chart elements/areas depending on the viewing context.

During the workshop UNH pointed out that the easiest structures to set up were those that were already in tables since each row would be the equivalent of an instantiated object and the columns would represent the components/attributes of that object. This helped to set the focus on tables that would be the priority; one of which was the bridges table.

While working with the bridges table, which was thought to be a “low hanging fruit” (meaning possibly able to begin to implement the new system starting with the bridges structure), and comparing it with the ENC database (based on the S-57 data) and the S101 specification, it was obvious most of the information for a bridge object was already in the ENC. It began a discussion of the following:

1. how to connect to the database to use the information already there
2. harmonization between the data sources
3. any elements for bridges that might extend the S-101 data structure.

It was noted that the bridges table did have an attribute that the ENC/S-101 data did not:

Span fixed vertical clearance lat/lng point

It was agreed upon to keep checking against the ENC/S-101 data structure as well as other related structures (S-104/water levels, S-111/surface currents, S-123 Radio Services, S-124 Navigation Services, S-126 physical environment, S-412 weather overlays) as the data structures in the CP are created. Doing this will help to harmonize the CP data with any existing/proposed S-100 products and possibly see where/if new products might need to be created.

Outside of the table data, the data structure creation, based on textual paragraphs, has a few iterative steps as follows:

1. Find all the atomic elements (aka simple attributes)
   These would be easily parsable items like distances, weight, time of day, vessel class, month, etc.
2. Identifying the main subject/theme of the paragraph and creating tags for that theme around the entire paragraph.
3. Normalizing the tags used (making tag names consistent for each subject/theme)
4. Refining the details within the structures (creating more complex attributes)
5. Creating feature associations
6. Normalizing the feature associations

As each step and grouping of steps are iterated on and the data structures begin to form (being checked against the S-100 equivalents) the data will then be in a position to be harmonized with other datasets.

Aside from having a mechanism for harmonizing the initial data, it seemed obvious that the CP data that did overlap with the ENC should also have in place a mechanism for updating the future data. Ideally, the CP would just connect to the ENC database to populate the information it needed to ensure the data reflected the same values.

The discussion also led to a question of possibly redefining what the CP would consist of. Would it be a one stop shop for all supporting data to the ENC where it collates information from external sources? Or possibly a restructuring the CP into different sections/products: one a mariners type of almanac or encyclopaedia of static/infrequently changing information, and the other one that focuses on combining other data with new structures contained only within the CP made up of more changeable/necessary data.

During the data linking discussion our attention was also tuned to some issues regarding connecting the databases to each other and how one dataset could reference objects in another using unique identifiers.

Maritime Resource Names (MRN)
Coinciding with a recent paper that came out for review in the NIPWG (Draft guidance for managing the IHO Maritime Resource Names) there was a very detailed discussion about how the IHO MRN could be implemented and used to tie together the different datasets.

A consensus for exactly how to build the MRN among the US agencies will need to be made before the OCS would begin to implement the use of the MRN; however, the details of the discussion on building the address of the MRN led to the possibility of using an *ontology* that coincides with the Chart 1 categories. Since this is a commonly used and known hierarchy already, it would be something that the maritime domain would benefit by reusing.

The biggest problem that arose with trying to tie the CP data with the ENC data was the fact that the ENC data didn't have a unique *persistent* identifier. Since OCS uses an ESRI arcGIS interface the system automatically creates a unique identifier but the object could be deleted or divided and new identifiers would be created even though the real-world object is still the same. The use case scenario again dealt with bridges and how they are formed within the current system with multiple lines or polygons.

Internally, the best solution was to create a table of features/items that would need to be referenced from external sources (like a listing of items in the ENC that the CP references), this will also be the foundation for the CP database to have a more feature-centric structure. This table (one-to-one type) will also have an MRN that is persistent and unique for each item. Then another table (one-to-many type) will be built that will reference the global identifier in the ENC database and assign it to the MRN, allowing for the possibility of many ENC global identifiers for each MRN.

Figure 2. a solution for linking the MRN with each data type, CP_Features is a one-to-one table listing all features in the CP, CP2ENC a one-to-many table to link all the possible changes to ENC data to one CP feature.

In order to keep track of the changes for the object on the ENC a system would be put into place that creates a bounding box buffer zone around each ENC item (bridges in this case). Using the resulting polygon, a hit test would be performed with the lat/lng point of the CP bridge data. This would assign the appropriate ENC bridge parts to the CP bridge MRN and would be run at regular intervals to ensure all updates are recorded.

Another easier and possibly less intensive way to do this was to use the OBJNAM from the ENC categories within the Cultural Features layer and match them with the common names used within the Coast Pilot. Although there could be some ambiguities if there is more than one object with the same name only in those cases would a polygon type hit test need to be performed.

Conclusions
The current publication-centric production system of the US Coast Pilot is attempting to move from a publication-centric paradigm to a data-centric paradigm. To accomplish this, features within the CP (or any publication) associated with the ENC need to be linked up with a persistent unique identifier (IHO MRN). Features also need a context or a topic (using the well-known and established ontology from Chart One as a foundation) in order to collate similar information to aid in filtering the data (for assisting the user with information overload issues as well as to generate various S-100 products). Lastly, the topics need to have their own data structures to determine necessary attributes and associations between other structures and to help to add to or extend the S-100 product suit.

Recommendations
It is recommended that other HO's share their initial evaluations and research into implementation strategies moving from older systems to new systems to accommodate the S-100 standards.

Justification and Impacts

Note: FOR REASONS OF ECONOMY, DELEGATES ARE KINDLY REQUESTED TO BRING THEIR OWN COPIES OF THE DOCUMENTS TO THE MEETING

There are many reasons why this new data-centric paradigm for maritime publications would be beneficial to the producers as well as the mariners in this age of technology and customized data/interfaces including:

- Helping to reduce cognitive overload with ability to filter the textual data according to tasks/topics
- Assisting in understanding the chart better/more intuitively with data geo-referenced
- Minimizing the potential clutter on a nautical chart by not adding more symbols/words but possibly just adding weighted values to feature importance.
- Many more options for how the data is presented.
- Data transfer to other entities will be highly machine readable as well as filterable to reduce data size.

Switching to new systems/paradigms after years of establishing a functioning system is difficult to effectuate. It will take time to implement, to gain permissions and access to the ENC database, plan detailed steps and find the man hours to do it all. An extra cost comes in the form of hiring an outside contracting firm that will take over the development of the system. It is estimated the deliverables for the first phase to be complete within a year's time. The projected deliverables would include:

- A mechanism for identifying all geo-referenced features in the CP and collating them into a database table that contains a MRN and location/polygon bounds for that feature.
- An interface to allow a human-in-the-loop to create or verify automatically generated feature polygons.
- Access from NSD to the ENC database will be granted and used for a hit test to determine feature object within the ENC that coincide with the CP features, which will then store feature id's from the ENC into the common CP features to ENC features table.
- A table to represent the elements and hierarchy of the Chart One publication to enable machine readable chart one data files as well as to assist in categorizing the CP data/features.
- A convention for automatically generating persistent ID's that will be used as link to any new identifying system in the future, e.g. MRN's (once the IHO has stabilized it's recommended structure and uses)

OCS feels the financial and time impacts are worth the investment as the vision of the future of the nautical publications is now clearer than ever.

Action Required of NIPWG

The SNPWG is invited to:

a. support the initiative to share information on implementing S-100 nautical provision standards

b. take into consideration the listed issues involving the MRN implementation as they move forward solidifying the terms on how to generate and use them

c. note the work NOAA/UNH has done in support of the S-100 nautical provision standards and their integration into a working production system.