CTD-data inom Datavärdskapet för Oceanografi och Marinbiologi

Diarienummer HaV: 2730-17

Diarienummer SMHI: 2017/2078/107

Örjan Bäck

Martin Hansson

Lisa Sundqvist

SMHI Oceanografiska enheten, 2017-01-18

# Bakgrund

## Syftet med uppdraget är att utveckla förbättrad datahantering för att öka bevarande av och tillgänglighet till data enligt grundprinciperna kring öppna data och den strategi för miljödatahantering som såväl SMHI som Havs- och vattenmyndigheten uttalat stöd för.

## I datavärdskapet Oceanografi och marinbiologi ingår i dag inte att ta hand om, lagra och publicera högupplösta CTD-data utan endast CTD-data med låg vertikal upplösning (≥ var 5:e meter). Med CTD-data avses data från mätningar med CTD-sonder. Data som genereras är bl.a. konduktivitet, temperatur och tryck och resultatet ger vertikala profiler genom vattnet som visar salthalt, temperatur, vattendensitet och eventuellt ytterligare data från extra sensorer på sonden (exempelvis fluorescens och syrehalt).

## Mätningar görs med hög vertikal upplösning och det finns behov av att lagra och göra även dessa högupplösta data tillgängliga. SMHI har idag en intern filtjänst där all högupplöst CTD-data som SMHI själva samlar in lagras tillsammans med CTD-data från fiskeexpeditioner och övriga dataleverantörer som levererar högupplöst CTD-data.

## Nedanstående behov av högupplöst CTD-data från miljöövervakningen har identifierats.

## **Dataanvändare** efterfrågar ofta högupplöst CTD-data. CTD-data används som stödparameter vid utvärderingar av övrig miljöövervakning såsom närsalter, fiskerihydrografi, växt- och djurplankton, koppling till åtgärder (direktiv) och statusklassning, modellvalidering, följa inflöden i Östersjön, etc.

## **Kvalitetskontroll** förenklas avsevärt om högupplösta CTD-data finns tillgängligt. I de data som publiceras i dag kan det vara 25-50 meter mellan diskreta djup och om detta sammanfaller med en skiktning påverkar det övriga vattenanalyser stort. Högupplöst CTD-data är till stort hjälp eftersom det är den data som man kan använda för att se skiktningarna i havet och uppdelningen mellan vattenmassor.

## **ICES** efterfrågar både låg- och högupplöst CTD-data vid internationell rapportering (HELCOM).

## Efterfrågad data bör göras tillgänglig på samma webbplats som övriga marina data, dvs i SHARKweb och via SHARKdata. Därför bör CTD-data ingå i datavärdskapet Oceanografi och marinbiologi. Årlig driftskostnad för CTD-data inom datavärdskapet uppskattas till storleksordningen 100 000 kr.

# Uppdragsbeskrivning

Uppdraget ska fokusera på parametrarna, tryck/djup, temperatur, salthalt, fluorescens och syre (beräknade parametrar densitet och ljudhastighet). Det är dock viktigt att även data från eventuella andra sonder i samma CTD-kast hanteras inom datavärdskapet så att inga data kastas bort. Även dessa data ingår alltså i uppdraget. De olika data som samlas in vid ett CTD-kast kallas härefter gemensamt för CTD-data.

CTD-datas metadata ska databasläggas och själva CTD-data ska överföras till ett standard-filformat. Genom uppdraget ska SMHI göra det möjligt för användare att via SHARKweb söka ut data utifrån metadata och sedan få data levererad som separata filer eller zippade paket med filer som alla har samma format. Denna lösning öppnar även för möjligheten att i framtiden kunna ta hand om andra typer av datafiler i och med att bara metadata databasläggs medan datafilen ligger utanför sökdatabasen med sitt webbgränssnitt.

Avgränsningar: Endast CTD-data med tillhörande mätningar som insamlas för nationell miljöövervakning ingår, dvs. CTD-data från SMHI, SU, UMF och SLU. SMHI upprättar kontakt med dessa dataleverantörer för att efterfråga leverans av deras högupplösta CTD-data. Data som insamlats fr.o.m. 2016 ingår.

**Arbetsmoment:**

* Kartläggning av internationella metadataformat och standard-filformat för CTD-data. En kort skriftlig sammanfattning ska redovisas i slutrapporten för detta uppdrag. 40 h
* Fastställa metadataformat och standardformat för CTD-data. I arbetet ska det tas höjd för att arbetssättet även ska kunna fungera för andra format på sensordata. T.ex. gliders, undulerande provtagning, argo floats, ferrybox, m.m. 85 h
* Inleda utvecklingen och implementeringen av rutiner för kvalitetskontroll av CTD-data. Kvalitetskontroll på metadata samt kontroll av levererad data för parametrarna i fokus behövs för att säkerställa användbarheten av data. Rutiner för att granska och flagga data, exempelvis orimliga värden, spikar och instabila profiler kommer hanteras. 40 h
* Skapa dokument som sammanfattar utförda moment. Används internt på HaV och SMHI. 5 h
* Projektledning. 10 h

**Totalt tidåtgång**: 180 timmar motsvarande en kostnad av 162 000 kronor.

*Detta är den första delen i uppdraget att tillgängliggöra CTD-data. Uppdragstagaren kan inte avsätta mer resurser för detta arbete i år. Ytterligare arbetsinsatser kommer att krävas innan högupplöst CTD-data tillgängliggjorts inom datavärdskapet för oceanografi och marinbiologi.*

# Kartläggning av internationella CTD-data format

I arbetet med att ta fram en metadata-mall som ska användas vid leverans av CTD-data som ska ingå i datavärdskapet har en omvärldsanalys genomförts. Fokus har legat på att undersöka hur stora internationella datacenter och andra metadatakataloger hanterar metadata kring CTD data. Nedan redogörs för vilka organisationer som har beaktats och länkar till de webbplatser där metadata information har hittats redovisas.

Information har samlats in från det Britiska datacentret British Oceanographic Data Centre (BODC):

<https://www.bodc.ac.uk/resources/delivery_formats/odv_format/>

<https://www.bodc.ac.uk/submit_data/submission_guidelines/ctd_data/>

Det Amerikanska datacentret National Centers for Environmental Information (NODC) som ligger under National Oceanic and Atmospheric Administration (NOAA):

<https://www.nodc.noaa.gov/submit/submit-guide.html>

<https://www.nodc.noaa.gov/General/NODC-Archive/f022.html>

<https://www.ecofoci.noaa.gov/dataInfo/efoci_dataInfo.shtml>

Den Australiensiska geodata-katalogen Geoscience Australia:

<https://ecat.ga.gov.au/geonetwork/srv/eng/search#!a05f7892-eae4-7506-e044-00144fdd4fa6>

Metadatakatalogen Catalogue of Australian Antarctic and Subantarctic Metadata (CAASM):

<https://gcmd.nasa.gov/KeywordSearch/Metadata.do?Portal=amd_au&MetadataView=Full&MetadataType=0&KeywordPath=&OrigMetadataNode=AADC&EntryId=au0806_CASO_CTD>

Samt från Geonetwork:

<https://dc.niwa.co.nz/boi_dc/srv/en/metadata.show?uuid=36f7d9d6-60f3-45d5-9350-0fd333b5f82f>

Alla metadatafält som hittats på följande adresser har sammanställt tillsammans med de fält som SMHI har tagit fram som förslag för eventuellt kommande datavärdskap för video och bild. Ett Excel-dokument togs fram där liknande metadatafält samlades på samma rad och unika fält listades i kolumner. Totalt resulterade sammanställningen i runt 300 metadatafält. Den framtagna listan användes sedan för att ta fram det aktuella förslaget på metadatamall som redovisas i denna rapport.

# Inventering av format inom den nationella provtagningen

För att bilda sig en uppfattning av vilken typ av CTD-datafiler som finns idag inom den nationella provtagningen och som så småningom skulle behöva tas om hand om så gjordes en mindre inventering av tillgängliga instrument och variationer av filformat.

De tre nationella utförarna inom utsjöövervakningen använder idag olika CTD-modeller för sin övervakning. SMHI och Umeå marina forskningscentrum (UMF) använder i huvudsak modeller från Seabird, medan Stockholm Universitet (SU) använder en CTD från Sea & Sun. SU har en CTD från Seabird på sin nya båt R/V Electra, dock används denna i dagsläget endast till forskning och inte miljöövervakning.

SMHI: 2st Seabird 911+

1st Seabird 19+

UMF 2st Seabird 19+ / Seabird 19+ V2

1st Seabird 911+ (Används ej efter att KBV05 togs ut bruk men planeras tas i bruk efter att KBV181 byggts om)

SU 1st Sea & Sun modell?

1st Seabird 911+ (Används dock endast till forskning vid Östersjöcentrum)

CTD-data från Seabird har ett standardiserat dataformat (\*.cnv) som innehåller mycket metadata, kaliberingsdata, processeringsinfo samt mätdata. Dock saknar formatet kvalitetsflaggor och metadatan är i vissa fall bristfällig. Datafiler från Sea & Sun finns i någon form av textformat, SU lagrar även filer i Excel, dessa innehåller mycket begränsad metadata. Alla CTD-filer är ändå uppbyggda på liknande sätt, kolumnformat med viss headerinformation som gör att de går att hantera någorlunda smidigt. De undersökta varianter av filformat utgör inga hinder med planerade standardformat som presenteras nedan.

# Förslag till CTD-dataformat

Vid omvärdsanalysen framgick att många av metadatafälten har varit generella och inte anpassade speciellt för CTD-data. Metadata och data har dessutom ofta varit uppdelat med metadata i en fil eller på en sida online och data i en annan fil. Datafilerna har generellt varit väldigt enkla och inte innehållit mycket metadata. Med detta som grund så beslutades att ta fram ett förslag på ett nytt eget standardformat där filerna innehåller både data med kvalitetsflaggor samt all relevant tillhörande metadata. Vi övervägde även alternativet att lagra data i en fil och metadata i en annan men såg större fördelar med att ha allt samlat i en fil så länge den kan konstrueras på ett användarvänligt vis.

Högsta prioritet i förslag till nytt standardiserat format för CTD har varit att införa kvalitetsflaggor som saknas i de flesta befintliga CTD-dataformat samt att standardisera metadata. Detta görs genom att flaggkolumner införs i ett enkelt kolumnformat för data och metadata levereras in enligt en leveransmall som påminner mycket om leveransmallarna som finns för andra datatyper inom datavärdskapet.

Dataformatet kommer innehålla en övre metadatadel och en nedre datadel, den övre metadatadelen inleds med en standardiserad metadatadel med information från leveransmallen, se medföljande exempelfil, sedan följer metadata om sensorer som sitter på CTDn, detta är från en separat flik i leveransmallen, sedan följer ett block med information i form av text från en egen flik i leveransmallen; här är det tänkt att dataleverantören skriver lite information om datahantering, dataprocessering, kvalitetskontroller, instrumenthantering, kalibreringsrutiner och så vidare. Nästa del är metadata från den inlevererade CTD-datafilen, vissa format har mycket metadata, tex Seabird, medan andra har lite eller ingen metadata, men för att inte tappa bort någon info tas denna information med i sin helhet även i standardformatet, eftersom denna del varierar så är det helt fritt med varierande antal rader. Sedan följer datadelen med första raden en så kallad ”header” med kolumnrubriker med parameternamn som även återfinns i metadatadelen ovan med sensorinformation, på detta sätt går det att koppla en parameter till vilken sensor som genererat datan.

För att formatet ska passa så många användare som möjligt så har vi identifierat tre grupper med typer av användare; avancerad, bas, novis där vi tänker at den avancerade användaren använder sig av programmering för at maskinellt läsa in data för processering, visualisering och analys. Därför gör vi formatet maskinellt läsvänligt genom att lägga in ”taggar” i början av alla metadatablock i den övre delen, tex METADATA, SENSORINFO osv, se exempelfil. För användare på basnivå gör vi formatet ODV-vänligt, ODV står för Ocean Data View och är i första hand ett program för visualisering av marin data. Formatet ska enkelt kunna importeras i ODV för användare som inte i försa hand vill läsa in data med hjälp av programmering. För att formatet ska passa för ODV inleds alla rader i metadatadelen med //, vilket är ODVs kommenteringsmärkning, dvs dessa rader läses inte av programmet, vi har även med några inledande kolumner metadata i datadelen, detta är ODVs minimikrav på metadata och gör att filerna enkelt kan läsas och data kan visualiseras av ODV. För novis-användaren så gör vi formatet enkelt att öppna och använda i Excel, Metadatadelen blir visuellt tydligt uppmärkt och datadelen är tabbseparerad vilket Excel hanterar bra.

Tänkt dataflöde inom datavärdskapet:

1. Leverantör fyller i leveransmall:

* en del med metadata per kast/serie liknande som liknande provtagning, tex ”fysik/kemi”
* en del med instrument/sensor-information, påminner om analysinfo som finns i ”fysik/kemi”-mallen, kräver generellt en lite större initial insats som sedan borde gälla för mycket data (det är ganska sällan som sensoruppsättningen på en CTD ändras).
* en del med information i textform där data och instrumenthantering och processring m.m. beskrivs.

1. Leverantören skickar in mallen och tillhörade CTD-datafiler
2. Datavärden läser in information från mall och CTD-datafiler och genererar nya datafiler i standardformat, samtidigt automatgranskas data. Visuell kvalitetskontroll görs även som en del av granskningen. Eventuella misstänkta värden stäms av med leverantör och rättas eller flaggas innan filerna läggs upp på sharkdata.se och metadata läses in och görs sökbara i sharkweb.se.

# Fortsatt arbete

Förslaget av standardformat och leveransmall behöver presenteras och granskas av fler för att kunna få in synpunkter på hur användarvänligt det upplevs och om någon metadata missats.

Kod för automatiska kvalitetskontroller har endast kunnat påbörjas, inte till lika stor del som vi först hade planerat och hoppats hinna med, så där återstår lite mer arbete än planerat till del 2 av projektet.

Kod för automatiserad hantering och formatering från mall och datafil till standardformat är enligt plan med i del 2 av projektet, men det har självklart funnits i tankarna när vi designat leveransmall då den är gjord för att kunna automatiskt läsas och konverteras till standardformatet tillsammans med tillhörande datafiler.

På SMHI utvecklas nu ett verktyg för hantering av Ferryboxdata, går under arbetsnamnet GISMO, Gadget for In-Situ Measurements Oceanography. Tanken är att verktyget ska användas för granskning, analys och visualisering av data och det finns även planer på att utöka GISMO till att klara data från alla typer av mätinstrument, till exempel CTD. Troligen kommer SMHI att försöka lägga ovanstående kod för granskning under GISMO, men eventuellt skulle även koden för hanteringen av datafiler kunna ligga där. GISMO kommer att släppas som ett öppet verktyg tillgängligt för alla, detta skulle då kunna bli ett komplement till ODV, eller för den delen egenskriven kod, för hantering och visualisering av CTD-data.