2. Data Science Symposium

ABSTRACTS

Lectures and Posters on Thursday, 6.12.2018

13.00 Uhr
Hans Pfeiffenberger, AWI
Welcome, Data Science @AWI, GEOMAR, ... Helmholtz
Short intro to and recap on the emerging "tradition" of data science symposia at Helmholtz Centers (in the research area of Earth and Environment) and the scope of Helmholtz initiatives in the area of data (science) in general.

13.15 Uhr
Marlene Klockmann HZG
Reconstructing past North Atlantic variability within the Reduced Complexity Models (REDMOD) project

Marlene Klockmann, Eduardo Zorita and Corinna Schrumm, Helmholtz-Zentrum Geesthacht, Germany

The REDMOD project is a pilot project from the Helmholtz Incubator initiative. It aims at developing surrogates of complex and computationally expensive numerical models, quantifying uncertainties and discovering lower-dimensional subspaces in high-dimensional systems. It combines researchers from different disciplines such as plasma physics, health sciences and climate research, and thus fosters the exchange of existing methods and the development of novel approaches across disciplines.The main objective of the HZG contribution is the application of data-science methods to combine the information of long paleoclimate simulations with proxy data (e.g. tree rings, lake sediments, ice cores) to provide a reconstruction of past climate states. Ultimately, we aim at obtaining spatially complete climate field reconstructions which can help to reduce the uncertainties inherent in simulations and proxies. As a starting point, we are testing new approaches for the reconstruction of the North Atlantic Multi-decadal Variability based on Gaussian Process Regression (GPR). We are investigating whether the non-linear GPR approach can capture the magnitude and extremes of the past variability better than traditional linear regression methods, which are known to underestimate past variability.

13.30 Uhr
Lars Nerger
Enhancing data sets through data assimilation
Data assimilation combines observational data with numerical simulation models. The methodology allos to improve the initialization of model predictions, determining model deficiencies, but also to enhance data sets by augmenting the data with dynamical information from numerical models simulating e.g. ocean physics or biogeochemistry. This combination can fill data gaps by an interpolation which accounts for the dynamical information provided by the numerical model. Further the observed information can be used to improve unobserved variables, and even fluxes. This is accomplished through the use of dynamically estimated cross-covariances between the observed and unobserved variables. The assimilation can result in data sets which, at the resolution of the model, exhibit smaller errors than using the observations or the model alone. I will discuss the method of ensemble-based data assimilation on the example of ocean-biogoechemical modeling with the assimilation of satellite ocean color data.

13.50 Uhr
Dieter Piepenburg
Marine biodiversity research in the Arctic, North Sea and Antarctic: scales, exploration, sharing, and modelling.
Dieter Piepenburg, Casper Kraan, Jennifer Dannheim, Katharina Teschke, Kerstin Jerosch, Paul Kloss, Hendrik Pehlke & Tom Brey
Worldwide, marine biotas are affected profoundly and at large scales by accelerating environmental shifts caused by climate change. Moreover, increasing human exploitation of ocean resources add further cumulative pressures on marine ecosystems. Substantial impacts on biodiversity, ecosystem functions and services are expected. To develop evidence-based governance strategies for mitigation and adaptation to ensure long-term conservation and sustainable use of the oceans, seas and marine resources (UN SDG 14), it is critical to understand and predict the profound ecological consequences of these regime shifts. The relationships between environmental drivers and ecosystem functions have to be identified and analyzed from local to regional to ocean-wide scales in order to reach these overarching objectives. We address this challenge by means of a cross-latitude knowledge system on marine biotas. Underpinned by international efforts to combine data and expertise, this platform integrates quality-controlled and geo-referenced data on marine biotas from the Arctic, North Sea and Antarctic in a public knowledge system. It allows to (a) provide ecological baseline-data to gauge marine ecosystem changes, (b) couple environmental drivers and ecosystem functions/services on multiple spatio-temporal scales, (c) model current and future ecosystem scenarios in response to external forcing, and (d) create online stakeholder-oriented visualization and analysis tools. The talk will demonstrate the huge benefits of up-scaling marine biotic data with our system, our achievements to support data management, exploration and sharing, as well as first results of community-level distribution models to discern marine biotas on multiple scale and in relation to multiple-factor environmental forcing.

14.10 Uhr
Barbara Niehoff
Ecological information from underwater imaging: From raw images to a data base on zooplankton abundances in the Arctic
The plankton recorder LOKI provides high-resolution pictures, continuously taken by a 6 Megapixel camera during vertical hauls from 1000 depth to the surface. Linked to each picture, hydrographical parameters, e.g. depth, salinity, temperature, oxygen concentration and fluorescence, are being recorded. This allows to exactly identify distribution patterns in relation to environmental conditions. During one cast, up to 40,000 pictures are recorded. To check the quality of the images and to annotate each with taxonomical information, we have established a workflow, using different computer programs. To perform meta-analyses, all image and metadata obtained during various cruises are included in a postgreSQL data base which now allows to easily extract information on specific species and taxa.

14.25 h and 17.20 h
Poster and Live-Demos

Stephan Lange
Current methods and data structure and analysis in permafrost observatories and MOSES

Axel Behrendt
UDASH - Unified Database for Arctic and Subarctic Hydrography
Oceanographic data in high latitudes are sparse in both space and time. Most of these data are publicly available from different online archives. They often contain redundant profiles and data of different quality. To date, none of these archives offers a complete collection of all available temperature and salinity (T/S) measurements in the Arctic Ocean with a uniform quality level. We therefore compiled UDASH, a comprehensive hydrographic database of the Arctic Ocean, which aims at including all publicly available data. It so far consists of 288 532 quality-checked oceanographic profiles between 1980 and 2015, starting at 65°N.

Gregor Pfalz
Data Science at the „Polar Terrestrial Environmental Systems"-Group at AWI
Author: AWI Polar Terrestrial Environmental Systems
Within the working group "Polar Terrestrial Environmental Systems" at AWI many strikingly different but growing environmental data collections are used and established. These comprise vegetation, paleontological, bio- and geochemical and sedimentological data. Research-level data are collected at expeditions and generated in the laboratories for Paleogenetics, Micropaleontology and stable Isotopes. Community-level data collections are assembled from various publicly available databases that cover northern circumpolar regions up to the whole northern hemisphere. Special focus lies on areas with overall low data availability i.e. data collected for the Siberian treeline zone and for the Tibetan Plateau.
Research-level data collections: During various expeditions, sediment cores and extensive datasets on vegetation were collected in tundra, taiga and the Siberian treeline ecotone. Data like vegetation coverage or tree height and age help to identify spatio-temporal changes within the treeline ecotone. Modern and ancient DNA samples are analyzed to unravel changes in taxon composition to genetic variation on population level (e.g. for diatoms and Pinaceae) trough time. For example, this research aims at an in-depth understanding of tree line dynamics of different larch species. For paleoclimate and -vegetation reconstruction, own pollen datasets as well as a combination of available datasets for the whole northern hemisphere are analyzed. Group members are data authors and data users.
Community-level data collections: We assemble data collections for different scientific domains

i) A digital pollen data collection of the Northern hemisphere is collected from data sets from repositories and data collections and will be taxonomically and temporally standardized to produce higher-level derivative data
ii) A digital data collection for ESA Climate Change Initiative CCI Permafrost focuses on permafrost temperature, active layer and vegetation data from Northern and Southern hemisphere permafrost regions. The data are collected from the WMO/GCOS Global Terrestrial Network for Permafrost (GTN-P) and further national and international published data collections and will become part of the CCI+ Climate Research Data Package (CRDP) in the ESA CCI Open Data Portal and will thus be publicly available also for the broader climate science community.
iii) A sediment core data collection with focus on the Russian Arctic (ARCLAKES): Supported by two PhD positions associated with the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS), we will establish a data analytics system to manage and standardize abiotic and biotic sediment core parameters and analyze ancient DNA that describe polar terrestrial environmental variability.

Bridging the gap between geo- and bioscientific data, proxy processes, statistics and physical climate models, the Earth System Diagnostics group develops a quantitative approach for the use of paleo-environmental observations to reconstruct climate and environmental variability.
A major endeavor is the standardization of the datasets to make data reusable in order to meet domain-relevant community standards. To comply with state of the art methods, we maintain close collaboration with Helmholtz data science initiatives and the Einstein Center Digital Future.

15 Uhr
Kay Emeis HZG

Helmholtz Coastal Data Center: Status and Plans
The Helmholtz Coastal Data Center (HCDC) will over the next 5 years create a digital infrastructure for data capture and curation (expeditions and laboratory analyses; near-real-time data of fixed and mobile observatories), integration of observational data with model results and large data products, such as remote sensing data sets (Model Analysis Tool), and information dissemination. HCDC will be part of a marine data hub within the national research data infrastructure and will provide data services for partners in the coastal science community. Core themes of the current phase 1 (year 1-3) are a) integration of COSYNA, coastMap, coastDat following national and international standards and agreements; b) establishment of interfaces and work flows with existing (PANGAEA, DKRZ) and emerging data structures (NFDI4Earth, Deutsche Allianz für Meeresforschung), and c) software development and GIS services. Phase 2 will build on this to design information portals and an infrastructure tentatively named KüstenCloud. Our presentation will focus on existing components and forthcoming developments.

15.30 Uhr
Roland Koppe
O2A-HDF - New Developments
We enable scientist to describe the metadata of platforms, devices and sensors, which is versioned and citable. Data can be streamed in near real-time or ingested in delayed mode into the data storage. Raw data as well as data products are archived and published in PANGAEA. – But how does it fit together and how to work with this data and metadata?

15.45 Uhr
Philipp Fischer
New in situ observation technologies in coastal ecology: a (data) challenge or a nightmare for many ecologists
In times of climate change and digitalization, also aquatic ecology significantly changes. While only few years ago, many "marine ecologists" mainly relayed on discrete samples or strictly hypothesis-based experiments in the lab or field, the tremendous progress in sensor development and sensor automation over the last years led to a kind of paradigm change. Today's sensor and data processing capacities are so well developed that classic experimental designs in the lab can often be perfectly complemented by field monitoring approached to test if hypothesis based on in vitro experiments reflect in situ conditions. Unfortunately, academic education in handling and analysing larger sensor data sets has not been adapted so fast. Data handling and visualization programs like "Excell et al." are rather the common standard than the exception in many ecological work groups. On the other hand, explicit data science service support for data handling and analysis is not very common in many research institutes. This leads to a significant gap between data availability and data use preventing the necessary step forward to an data-integrative research strategy in many ecological research areas.

16.05 Uhr
Willi Rath, Geomar
Python-based analyses of marine data on the cloud and on HPC
The first part will be a short introduction to the parallel-analytics tool Dask. Dask provides different ways of describing analytic work loads as directed acyclic graphs splitting the task at hand into many sub-tasks which can be evaluated on nearly arbitrary compute resources.
The possibilities range from larger-than memory calculations on a personal laptop (with long wall times) to massively parallel analyses with very short wall times and elastic allocation of resources that allow for interactive work with whole climate model runs. The second part will be a demonstration running an actual anlysis on a high-performance computer or on a cloud-based cluster.

16.25 Uhr
Sebastian Mieruch
webODV -- Ocean Data View online
Ocean Data View (ODV) is a widely used software package for the analysis, exploration and visualization of oceanographic and other environmental data. It plays a fundamental role in the SeaDataNet (SDN) community and is heavily used for data file and parameter aggregation as well as for data quality control and visualization. In the framework of the SeaDataCloud (SDC) project an online version of the ODV software is being developed

called webODV. The online webODV tool will provide typical ODV functionality as a collection of modular web services. In the background, the modular web services make use of the full power of ODV running on the server side. The basic concept is to provide a user-friendly Browser interface which communicates with ODV on the server. On the server we run a special version of ODV that is equipped with a secure WebSocket server and allows bi-directional communication with the client via encrypted WebSocket messages.
In this presentation we will show:

- the basic concept of webODV
- the integration into the SDC project
- a live demonstration of a data extraction and a quality control service
- plans and potentials of further webODV developments

16.45 Uhr
Rolf Wittig
Integrating database functionality into a wiki-system: the AWIki project
The basic idea for the AWIki project is to enhance a wiki-system by combining it with database functionality. A wiki-system (e.g. Wikipedia or Confluence) provides an intuitive and transparent interface whereas a database (e.g. MySQL) allows to store and access data in an exact and organized manner. The AWIki approach overcomes the disadvantages of the wiki-system (data is not parameterized) and the complicated database access (SQL). AWIki is implemented as a web application provided by a virtual server running a java package within apache tomcat and a MySQL database. The main goal is to provide a tool for everyday use in our section. To have all the information at one place allows a high degree of networking. The already large size of the applications is constantly growing and will be presented in detail and with examples.
- contact details (persons, companies, institutions),
- personal information (list of the absences),
- device management (inventory, booking, SAP), chemicals,
- license management (number, assignment),
- room management (links to devices or persons, also booking),
- event management (e.g. expedition),
- simple project management,
- experimental protocols,
- workflows, storyboards, checklists,
- store paper, talks, minutes, contracts and lists of all sorts,
- display up-to-date measured values,
- backup and archive data,
- manage risk assessment, safety instructions etc

17.05 Uhr

Tim Leefmann

UltraMassExplorer: online and open access evaluation for mass spectrometric analyses of complex organic matter

T. Leefmann, B. P. Koch, S. Frickenhaus, C. Schäfer-Neth, J. Matthes, A. Steinbach

Abstract

In the past 15 years, direct-infusion high-resolution mass spectrometry (HRMS) has become an important analytical tool in the chemical analysis of dissolved organic matter (DOM). With increasing sample throughput, the processing of HRMS datasets with thousands of mass peaks per spectra has become a bottle-neck in the analytical pipeline, and thus a major challenge for analytical chemists. Therefore, we developed and published an interactive web application named UltraMassExplorer (UME)[1,2] that allows for the fast and efficient analysis of large HRMS datasets. UME was developed using the *R* programming language and the *Shiny* package that allows non-specialists to create interactive web applications. Currently, the web application is deployed on a docker server allowing up to five external users to run analyses in parallel. As an alternative to the web application, we provide the source code of UME under the Affero General Public License (AGLP) via a gitlab repository[3] for downloading and running the application locally. In the past three weeks after publishing the application and the accepted version of the article that introduces our data pipeline, users from seven countries accessed and used the UME web application. Currently, we use the UME source code as a blueprint for designing another web application for handling the evaluation of large datasets recorded by 3D fluorescence spectroscopy. In future, we plan to deploy our web applications via the AWI marketplace that facilitates user authentication and control over computational resources.

2. Data Science Symposium

ABSTRACTS

Lectures and Posters on Friday, 7.12.2018

9.00 Uhr
Stephan Frickenhaus
Overview on DAM/ NFDI
The three marine research centers AWI, GEOMAR and HZG are defining a core of data services for implementing FAIR and OPEN principles. Chanelling and presenting marine research data for a bigger consortium NFDI4EARTH is planned. A pilot-Implementation will be realized in 2019/2020, coupling the institutes data management frameworks O2A (AWI), OSIS (GEOMAR) and HCDC (HZG) by techniques of standardized metadata harvesting, search functionality and transparent/ open file sharing.
In parallel a Data-IT will be initiated for Deutsche Allianz für Meeresforschung (DAM), setting up a data portal that incorporates data from the data flows from the four major research vessels in Germany (Maria S. Merian, Sonne, Polarstern, Meteor). A prerequisite is that the further participating university institutes build capacities to organize their data workflows from ship to storage servers in a comparable way, e.g. using DShip software and implementing "Landstation services", including quality control and data publishing (obtain DOI).

9.15 Uhr
Anne-Cathrin Wölfl, Geomar
Transit Bathymetry: Making the most of every nautical mile
Wölfl[1], Anne-Cathrin; Devey[1], Colin; Augustin, Nico
[1]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

Approx. 230 million km^2 of deep sea ocean floor still need to be mapped in high resolution. Research vessels that operate internationally usually have long transit routes, when going from one study area to the other. In addition, they transit more or less randomly across the oceans to reach remote targets. Often these routes lead them through international waters, where data recording is unrestricted. In order to use every possible opportunity to collect bathymetric information, three German and one Dutch research vessels collect multibeam data with the ship´s own echo-sounder systems on their transits through international waters and occasionally in national EEZs. After the data has been recorded, the ship operators send the raw data including metadata information and sound velocity profiles (if available) to the GEOMAR Helmholtz Centre for Ocean Research, Kiel, where the data is quality checked, processed and several data products created. The data is uploaded under a creative commons license to the international Pangaea Data Publisher for Earth & Environmental Science and to the Data Centre for Digital Bathymetry from the International Hydrographic Organisation (IHO-DCDB). Of key importance for getting the support of both scientists and crew for transit mapping has been our commitment to make the data freely and openly available for all.

9.45 Uhr
Stefan Hendricks
SIRAL & SMOSice: sea-ice thickness from remote-sensing for climate research and operational applications.
The variability and change of sea-ice thickness and volume is essential to understand Arctic change and central part of the scientific objective of the sea ice physics section. The requirements for a sea-ice thickness observing system that addresses these objectives cannot be met without satellite remote-sensing. However, meeting the accuracy requirements for sea-ice thickness from remote sensing is challenging and retrieval algorithms are currently less mature than for sea ice parameters such as sea ice concentration. Originally to support its own research, the sea ice physics section has invested in algorithm development for obtaining sea-ice thickness information from satellite radar altimetry. The AWI internal SIRAL (Sea Ice Radar Altimetry) project of the sea ice physics sections also works in close cooperation with the glaciology section that employs similar methods for land ice applications. Today, AWI is one of the leading institution for sea-ice thickness remote sensing worldwide and it contributes to all related European research programs, such as the ESA Climate Change on Sea Ice (CCI), the Copernicus Climate Change Services (C3S) and the Copernicus Marine Environmental Monitoring Service (CMEMS). In a recent addition, AWI now hosts the official ice processing chain (SMOSice) for data from the European SMOS satellite, which is highly complementary to the SIRAL activities. Therefore AWI has now a powerful in-house capability to generate and improve sea-ice thickness information for climate research and facilitate model assimilation and initialization for sea ice prediction efforts. The sea-ice thickness processing chains of SIRAL and SMOSice ingest calibrated sensor data of several satellites and derives a series of higher level data products with a current data volume of approximately 150 TB. Automatic data ingestion by the IT department keeps primary and auxiliary input data up-to-date. Since we are producing climate data records, one of our main priorities is the reproducibility and traceability of data products. Our software engineering and product generation is therefore aligned to these priorities and we employ strict version control and an open-source approach. We also comply with requirements for self-describing and quality controlled data formats according the Climate & Forecast (CF) conventions. From the winter season 2018/19 on, the processing chain is run automatically to support operational use of sea-ice thickness information from remote sensing to support international sea ice monitoring programs and to enable feasibility studies for use of sea-ice thickness data for numerical weather prediction.

10.05 Uhr
Veit Helm, Angelika Humbert, Nils Dörr, Steven Franke, Niklas Neckel
SAR Data processing of airborne ultra wide band radar and satellite altimetry
Applications of Remote Sensing techniques and data processing tools in the glaciology group of the Alfred-Wegener-Institute is wide spread. As we are studying processes in the upper layers of snow and firn, and the internal structure of the ice as well as the bed topography and bed conditions across the vast ice sheets of Antarctica and Greenland the data pool and sensors are manifold. To study ice thickness and internal structure of the ice sheets e.g. origin of shear margins of ice streams and ice stream genesis, the glaciology group purchased the ultra-wide band radar system (UWB) from CRESIS, University Kansas, in 2015. The powerful system, installed in Polar5/6, transmits coherent radar chirps within a frequency range of 30 to 520 MHz. 24 channels receive the signals enabling on the one hand to SAR focus the data along track to an unprecedented resolution of 1m and on the other hand to map the bed topography in two dimensions. This enables us to resolve internal channels within the ice, where supra- glacial water drains down to the bottom of the sheets, changing the ice flow characteristic's. Data amount and processing costs are high, as the system generates up to 40 Terra Bytes on data per flight campaign.
Another branch in our group uses satellite data to map the spatial and temporal variability of ice surface velocity of dedicated glaciers as well as ice sheet wide, using the Sentinel 1 and TerraSAR-X archive of ESA and DLR, respectively. We developed a full branch of software tools to use calibrated sensor data in a semi-automatic parallel processing to deliver higher level products, like filtered velocity fields, to be assimilated in the house modeling environment. Next to the velocity maps we deliver mass balance estimates of the ice sheets in monthly and multi-year resolution from altimetry data like CryoSat-2. In close connection with the sea ice group a processing line was developed to produce higher level products starting from calibrated sensors data provided by the space agencies. The data pool comprises of a handful of different altimetric sensors reaching back to the early nineties. Currently we are working on a 25 years long time series of elevation and mass change. Up to now we are using a total amount of 100 TB of satellite data including manifold products of our sensitivity analysis.

10.25 h and 13.05 h
Poster and Live-Demos

Stephan Lange
Current methods and data structure and analysis in permafrost observatories and MOSES

Axel Behrendt
UDASH - Unified Database for Arctic and Subarctic Hydrography
Oceanographic data in high latitudes are sparse in both space and time. Most of these data are publicly available from different online archives. They often contain redundant profiles and data of different quality. To date, none of these archives offers a complete collection of all available temperature and salinity (T/S) measurements in the Arctic Ocean with a uniform quality level. We therefore compiled UDASH, a comprehensive hydrographic database of the Arctic Ocean, which aims at including all publicly available data. It so far consists of 288 532 quality-checked oceanographic profiles between 1980 and 2015, starting at 65°N.

Gregor Pfalz
Data Science at the „Polar Terrestrial Environmental Systems"-Group at AWI
Author: AWI Polar Terrestrial Environmental Systems
Within the working group "Polar Terrestrial Environmental Systems" at AWI many strikingly different but growing environmental data collections are used and established. These comprise vegetation, paleontological, bio- and geochemical and sedimentological data. Research-level data are collected at expeditions and generated in the laboratories for Paleogenetics, Micropaleontology and stable Isotopes. Community-level data collections are assembled from various publicly available databases that cover northern circumpolar regions up to the whole northern hemisphere. Special focus lies on areas with overall low data availability i.e. data collected for the Siberian treeline zone and for the Tibetan Plateau.
Research-level data collections: During various expeditions, sediment cores and extensive datasets on vegetation were collected in tundra, taiga and the Siberian treeline ecotone. Data like vegetation coverage or tree height and age help to identify spatio-temporal changes within the treeline ecotone. Modern and ancient DNA samples are analysed to unravel changes in taxon composition to genetic variation on population level (e.g. for diatoms and Pinaceae) trough time. For example, this research aims at an in-depth understanding of tree line dynamics of different larch species. For paleoclimate and -vegetation reconstruction, own pollen datasets as well as a combination of available datasets for the whole northern hemisphere are analysed. Group members are data authors and data users.
Community-level data collections: We assemble data collections for different scientific domains

i) A digital pollen data collection of the Northern hemisphere is collected from data sets from repositories and data collections and will be taxonomically and temporally standardized to produce higher-level derivative data
ii) A digital data collection for ESA Climate Change Initiative CCI Permafrost focuses on permafrost temperature, active layer and vegetation data from Northern and Southern hemisphere permafrost regions. The data are collected from the WMO/GCOS Global Terrestrial Network for Permafrost (GTN-P) and further national and international published data collections and will become part of the CCI+ Climate Research Data Package (CRDP) in the ESA CCI Open Data Portal and will thus be publicly available also for the broader climate science community.
iii) A sediment core data collection with focus on the Russian Arctic (ARCLAKES): Supported by two PhD positions associated with the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS), we will establish a data analytics system to manage and standardize abiotic and biotic sediment core parameters and analyse ancient DNA that describe polar terrestrial environmental variability.

Bridging the gap between geo- and bioscientific data, proxy processes, statistics and physical climate models, the Earth System Diagnostics group develops a quantitative approach for the use of paleo-environmental observations to reconstruct climate and environmental variability.
A major endeavour is the standardisation of the datasets to make data reusable in order to meet domain-relevant community standards. To comply with state of the art methods, we maintain close collaboration with Helmholtz data science initiatives and the Einstein Center Digital Future.

11 Uhr
Mario Hoppema
GLODAPv2: a living data, quality-controlled, internally consistent database of oceanic carbon-relevant data.

Global Ocean Data Analysis Product version 2 (GLODAPv2) is a major synthesis effort involving the GLODAP Reference Group with scientists from all continents. In the framework of GLODAPv2, we have assembled data from some 800 cruises mostly from the WOCE, CLIVAR and GO-SHIP programs, where measurements of inorganic carbon (TCO2, total alkalinity, partial pressure of CO2, pH) and other carbon-relevant variables (oxygen, nutrients, CFCs) have been made. It has been constructed by merging three existing carbon data synthesis products, namely GLODAP, CARINA, PACIFICA, and adding data from around 200 new cruises from all oceans. The data included in GLODAPv2 have been subjected to primary quality control (checking for outliers) and subsequently been bias corrected: All deep ocean data collected at the various cruises were examined using a crossover analysis of deep data followed by a robust inversion routine. Adjustments to be applied were vetted only after rigorous checking by the group members. GLODAPv2 includes the data files from the data originators, the integrated bias-corrected data product, and additionally mapped climatologies of the variables total dissolved inorganic carbon (TCO2/DIC) and total alkalinity. GLODAPv2 is a sustained effort within the International Ocean Carbon Coordination Project (IOCCP). A full new data product (next: GLODAPv3) will be produced about every 10 years. Yearly or bi-yearly updates will be provided based the previous inversion data (e.g. GLODAPv2.2018). An automation effort for submission of new cruise data to GLODAP is underway.

11.20 Uhr
Pier Buttigieg
The INTERNAS process: creating machine-readable interfaces to multi-stakeholder knowledge
Data and information systems naturally fuel the generation of knowledge. However, the expanding role of machine-readable knowledge in the discovery and mobilisation of data is equally important in distributed and FAIR-aligned systems. In this vein, semantic web technologies combined with formal knowledge representation offer one route to add artificial intelligence to data science research and applications. Here, we summarise early work in the INTERNAS project, which fuses social science and knowledge representation approaches to support the transfer of INTERNational ASsessment outcomes - with high relevance to the Earth and environmental sciences - into recommendations for stakeholders within the German national context. Effective knowledge transfer across governmental and non-governmental bodies, research organisations, and industry is essential for the implementation of international policies, treaties, and directives in national contexts. Such implementations largely define how national research activities can interface with societal goals and interests. However, the nuanced understandings native to each national sector are rarely harmonised or even visible across stakeholder groups. This disjointedness inhibits the development of efficient knowledge, information, and data flows across and between sectors. This is especially true when stakeholders must contend with subject matter beyond the direct scope of their standard intra-sectorial activities. Through a series of stakeholder workshops, semi-structured interviews, and surveys, we are eliciting multi-sectorial knowledge on international assessments such as IPBES and the EU's Common Agricultural Policy. We then apply the best practices of the Open Biological and Biomedical Ontology (OBO) Foundry to convert this knowledge into coherent ontologies, which may be used as resources for distributed data systems. In doing so, we reuse and extend well-adopted reference ontologies such as the Environment Ontology (ENVO) and the Sustainable Development Goals Interface Ontology (SDGIO). In this contribution, we will present the INTERNAS methodology along with some early outcomes and perspectives on how these can be used by data scientists and developers.