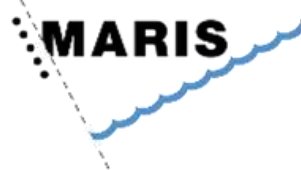


27-29 May 2024 



imdis

# International conference on Marine Data and Information Systems



# **Streamlining Data Submission: A Guide to Ensuring Data Quality and Compliance with ICES Data Screening Utility (DATSU)**

**Laura Andreea Petre**

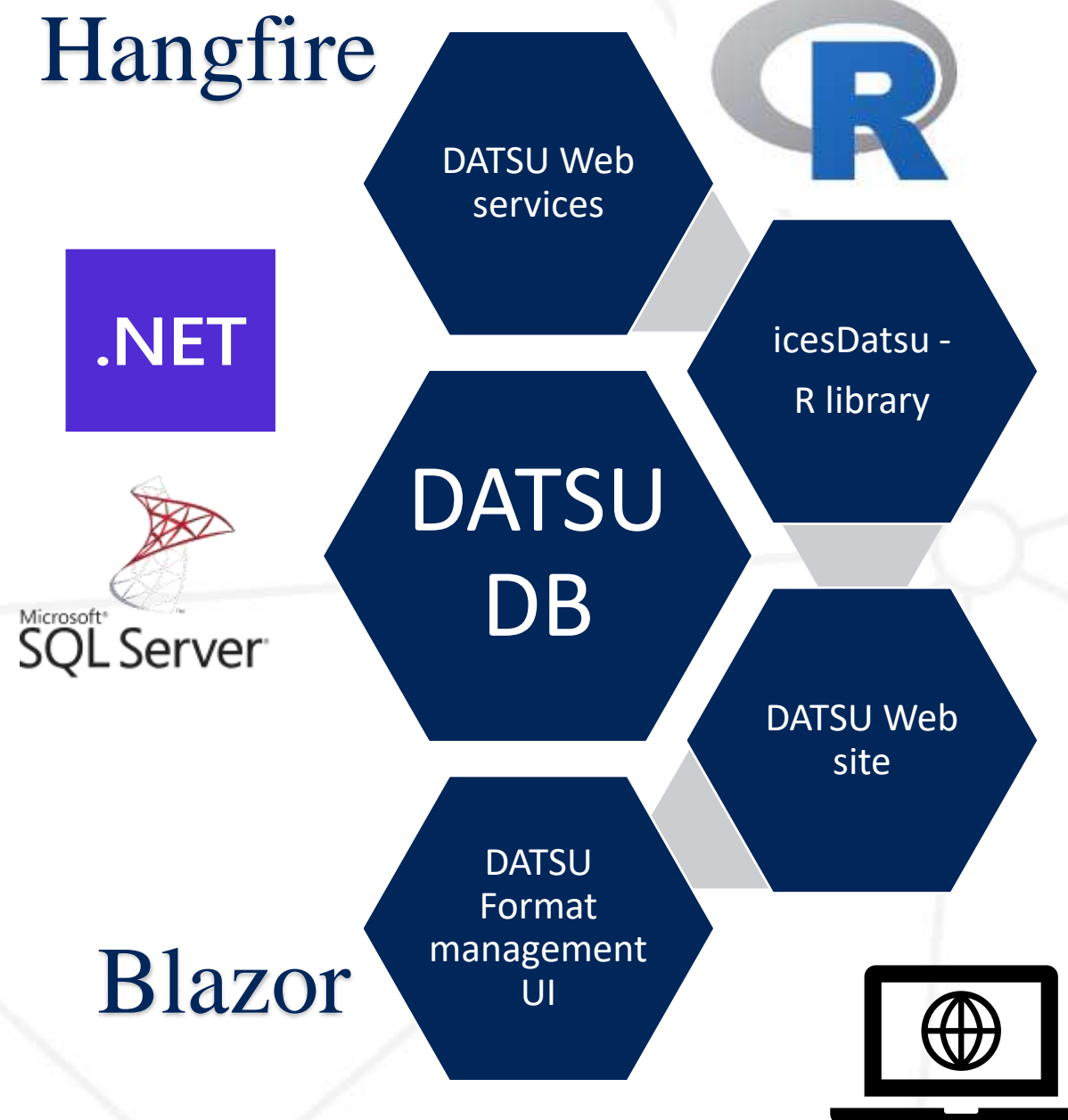
Carlos Pinto

Anna Osypchuk

Colin Millar

# Introduction to DATSU

- [DATSU](#) is the main data validation tool used by the ICES data centre.
- Designed to streamline and enhance data validation before uploading to ICES databases.
- Based on an SQL Server database (DATSU DB) with a user-friendly interface (DATSU UI).
- Utilizes web services for data submission and validation.



# History



**Early 2000s:** Initial development with key components: DATSU DB, ACCESS-based UI, and DATSU VB6 DLL.

**2007:** Extension to include SQL rules in screening reports.

**2016:** Reprogrammed to run as .NET web services with asynchronous runs for improved performance.

**2019:** Introduction of the R library (icesDATSU) for local data checks.

**2023:** UI migrated to a C# .NET and Blazor-based web application with enhanced functionalities.

# Components of DATSU



## DATSU DB:

- Stores format definitions and screened data.
- Central repository for data validation rules and vocabularies.

## DATSU Management User Interface (DATSU UI):

- Allows data managers to define data formats and quality checks.
- Provides an intuitive interface for managing data validation rules.

## DATSU API (webservice):

- Web Interface for data submission and validation.
- Can be called synchronously or asynchronously to check data files.
- Supports integration with other applications and programming environments.

## DATSU Web

- Allows data submitters to manually upload files for screening
- Allows viewing the screening report
- Exposes format information, data types, and vocabularies

## icesDATSU – R Library

- Provides functions to interact with DATSU web services.
- Allows experts to incorporate DATSU checks into their data pre-processing workflow.
- Supports local file checking before submission.



# Data validation and submission process



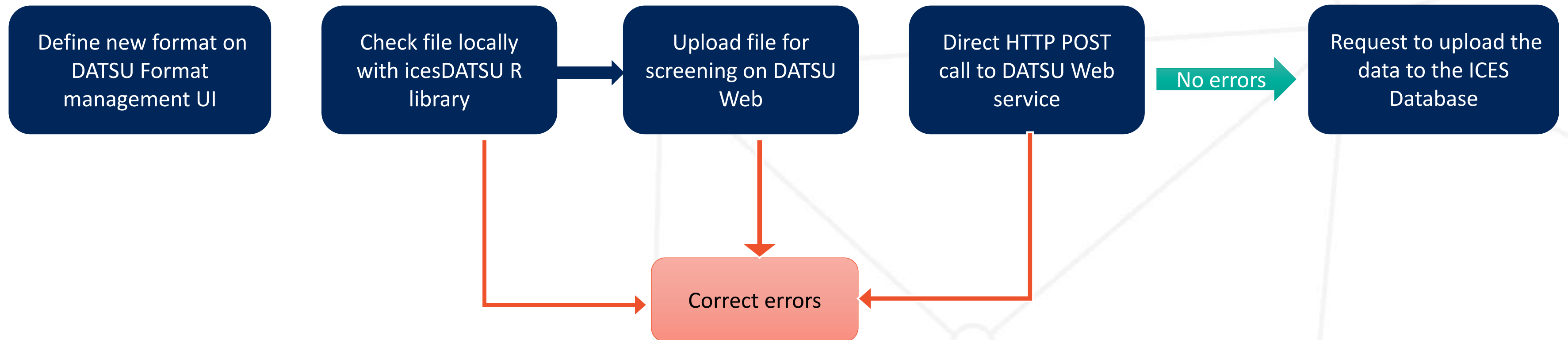
ICES Data managers



Data submitters



External server



# Data validation process cont.



**DATSU**

DATSU > List of operations > List of previous sessions

Name	UploadFile()
Description	This API of the web service screen file for errors and warnings.
Input	FileName EmailAddress DataSetVerID: for example DATRAS=59, Biota,CF=22 FileByteStream: IO stream of file
Output	SessionID: Successfully (Screening report is available by using SessionID), if SessionID is not available, the report is available by using the email address: <a href="mailto:accessions@ices.dk">accessions@ices.dk</a>
Example	Select file to be screened: <input type="button" value="Browse..."/> No file selected.  Select Data set version: <input type="text" value="Biodiversity - Cetaceans (JCDP)"/>  Select maximum number of errors: <input type="text" value="10000"/>  Enter e-mail: <input type="text"/>  <input type="checkbox"/> Send an email when session finishes? (if you check this option, you will receive an email when the session is finished) Note: By default DATSU will set the maximum amount of error messages to 5

Result On

**DATSU**

DATSU > List of operations > Screen a file

**Result from the ICES DATA Screening Utility program for the following data:**

Country: her.27.3a47d  
 Dataset: SAG Catch Scr  
 FileName: asd 2024 ns he  
 Email: xxx@ices.dk  
 Monitoring Year: 2025  
 Submitting date: 15/04/2024 14  
 Time to screen: 0.18 minutes  
 Number of records in file: 215  
 Max. Errors to return: 55000

Number of records per record type:	
record: AA	1 Row(s)
record: AK	16 Row(s)
record: AL	176 Row(s)
record: AN	12 Row(s)
record: AR	10 Row(s)

Number of Errors	Type	Reported Errors
2	error	Field value is invalid (Rel) - prevents entry to database

Record Line	Type	Reported errors
3	error	Field value is invalid (Rel) - prevents entry to database
3	error	Field value is invalid (Rel) - prevents entry to database

**DATSU**

The Data screening Utility (DATSU) Webservice provides the ability to query sessions and screen a file using a REST interface.

Each part of the web service is provided by a web service Application Programme Interface (API), which defines the name, input to, and output from the service for a particular data request.

The following web service APIs are provided:

Name	getDataverIDs
Description	This API allows the user to have a list of all the datasets IDs that can be screened in the DATSU API.
Output	The list of Datasets that can be screened in DATSU with the IDs
Example	<a href="#">Example to make a request to see the list</a>

Name	getRecordIDs
Description	A list of the available records for the specified format
Output	List of the supported records for the specified records
Details	<a href="#">Example of getting the list of the supported records for the ESAS format</a>

Name	getListQCChecks
Description	A list of the available QC checks for the dataset and/or record
Output	List of the supported available QC rules for that dataset and/or record, this list will include the source code of the QC rule in SQL
Details	<a href="#">Example of getting the list of QC rules for the VMS and logbook data</a> <a href="#">Example of getting the list of QC rules for the Seawater format but only for the record 91 (station record)</a>

Name	getDataFieldsDescription
Description	A list of the available fields for the specified format and record (optional filter).
Output	List of the supported records and fields, in this list it is specified the field type, if it is mandatory and if it is linked with any vocabulary.

International Council for the Exploration of the Sea (ICES) · Conseil International p

# Format management DATSU UI Web app



- Implemented using C# .NET and the Blazor framework.
- Focus on enhanced format management, improved user experience, and better validation for format changes.
- An automatic synchronization with the ICES Vocabulary Server was implemented using the Vocab Web API.
- Redundant features, such as species lists and reports, were omitted from the new version.

The screenshot displays the 'Version Details' page in the DATSU UI Web app. On the left, a sidebar contains navigation options: 'View datasets', 'Edit dataset details', 'View versions', 'View record types', and 'View field codes'. The main content area features a 'Back to Versions page' link and a 'Version Details' section. This section includes a table titled 'Datasets using version 1' with columns for Code, Description, Published, and Dataset details. Below this is a 'Fields defined' section with a search bar and an 'ADD FIELD' button. The 'Fields defined' table has columns for Field Code, DataType, Width, Heading, Detailed Description, and edit/delete icons.

Code	Description	Published	Dataset details
BITS	Baltic International Trawl Survey - before 2004	True	<a href="#">↗</a>
NS-IBTS	North Sea International Bottom Trawl Survey - before 2004	True	<a href="#">↗</a>
InterCatch	InterCatch 1. Version	True	<a href="#">↗</a>
IC_LogbookSurvey	InterCatch survey and logbook data	True	<a href="#">↗</a>

Field Code	DataType	Width	Heading	Detailed Description	
AgeLength	int	2			<a href="#">✎</a> <a href="#">🗑</a>
AgePrepMet	char	4	Age reading preparation method		<a href="#">✎</a> <a href="#">🗑</a>



# Benefits & limitations



- Exposes format information, data types, and vocabularies online and via web services.
- Allows for flexible data submission and screening methods.
- Generates comprehensive reports accessible through web pages and services.
- Shows the source code for quality checks.
- R package facilitates quick local data checks.
- Screening queue can be blocked by large files, causing delays.
- DATSU is heavily based on shared functionality and definitions for the format which can be both an advantage and a disadvantage.

# Future development

## Major change involving using message queues in DATSU

- Will avoid long waiting time and parallelize the process, by each dataset having its own queue and processor.
- In the future move to microservice architecture by separating the methods for running different types of checks.
- This will be implemented using RabbitMQ and MassTransit

## Functionality extension

- Allow the fields to be reported in another order in the text files if the user reports the headings.
- Accept JSON file format in submissions.
- Add logging for each change made using the DATSU UI by data managers.

Thank you for your attention.  
Any questions?

27-29 May 2024 



imdis

# International conference on Marine Data and Information Systems

