FIRMS Technical Working Group Meeting
Second Session
Rome, Italy, 01-04 April, 2008
WORD TO XML CONVERSION TOOL
Author: FSC Secretariat

Purpose of this document: This document describes the process to convert Word Document into FIRMS compliant XMLs using dedicated tools.

Revision history: This document is the second version of a document initially presented to FIRMS TWG1 (FIRMS TWG1/2005/Inf.3). Main modifications concern the update of the tool according to the upgrade of the FiMES schema. It includes a description of the Validator tool and describes how the tool can be extended to different partner needs.

Word to XML conversion Tool

Summary

1 Technical note

1.1 Purpose of this document

This document describes the process to convert Word Document into FIRMS compliant XMLs using dedicated tools.

1.2 Target audience

FIRMS partners

1.3 Scope

How to use the services and methods provided by this tool

1.4 Assumptions

Knowledge of XML and understanding of what XSL is. Knowledge of the Content management System

1.5 Technical needs

PC with Microsoft Word, an XML editor (XMLSpy, oXygen) and Upcast.

1.6 Definitions

A glossary is available in annex 3 for technical terms.

2 Preamble

2.1 Reasons and origin of this tool

The purpose of FIRMS is to provide access to a wide range of high-quality information on the global monitoring and management of fishery marine resources. To exchange information between partners and FIRMS and to drive its display, XML is the format used to store data.

The information content to be contributed in FIRMS (Aquatic resources, fisheries) can be provided by various tools.
The FIRMS partner having the responsibility to format and publish this information can use:

- One-Line Editing (OLE) if they are not willing to deal with the XML format. Using Web-based forms, OLE allows to do modifications on observations already loaded and to create new observations.
- Loading process that allows to load directly XML file related to an resource or fishery object. In that case, the full power of FIRMS is reached by the handling of the XML files.

To be loaded, these files need to be compliant with the current FiMES schema (version 3.6).

The tool to convert Word document into XML files was developed in order to provide an automatic conversion if the partner's template or source format made it possible. It is an efficient tool to generate upstream FIRMS XML files.
One of the FIRMS partner, ICES, handles as many as 150 stock status reports per year. To provide status and trends information to the FIRMS system in a timely basis for such a volume of information could only be envisaged if the task was semi to fully automated. At the beginning of 2005, a first mission in Denmark for ICES was realized to see how to implement a procedure to easily create a huge quantity of Stock summaries XML compliant with the FIRMS definitions from their original Word documents. A second mission in June 2007 allowed to develop and release full-features tools which generate automatically compliant XMLs from Word documents.

This tool is now used to generate XML files from NAFO scientific council reports. It is used as well internally in the fishery department of FAO for converting Word documents related to country profiles, aquaculture data into compliant XMLs which are then processed into the Fishery website of FAO.

2.2 Conditions and prerequisites

To be able to run this tool on a production basis, some conditions and prerequisites are necessary:

- Editing of original Word documents done according to strict Word template rules
- Internal skills available to intervene on the XML if necessary
- Arrangements made to ensure maintenance of the application.

Partners are mainly using Word documents for information production and data support.

In the case of ICES for example, these documents describing resources are called Stock summaries. These Word documents always have the same structure:

- Several paragraphs with one or two levels of headers with or without nested tables/figures.
- A list of tables/figures at the end of document.

The list of headers is very stable, some of the headings (and paragraphs) may be missing, but new ones will not appear. On that basis, a mapping between Word header and elements in the FiMES schema has been established. It signifies that a title present in the Word document is linked to an element available in the FiMES schema having an equivalent meaning.

Considering that the format of the reports provided by a partner respects these conditions, the tool can generate FIRMS compliant XMLs. Thereby, we will have one FIRMS XML generated per stock summary Word document.

In line with the Information Management Policy (IMP), the tool can consider the partner standards and classifications which are available in the Word documents and process them in order to make them available in the XML file together with the FIRMS equivalent standard terms.
It has been applied for the time being for ICES on the status and trends controlled terms (Exploitation rate and abundance level) for which a correspondence between ICES classification and FIRMS classification for stock status has been established.
It means that each XML for this partner will contain both classifications: Partner classification for display and data integrity purposes and FIRMS classification for internal data management purpose.

Moreover, the converter tool can be easily adapted to other partner needs as soon as a strict mapping has been established between standards partner titles and FiMES elements.

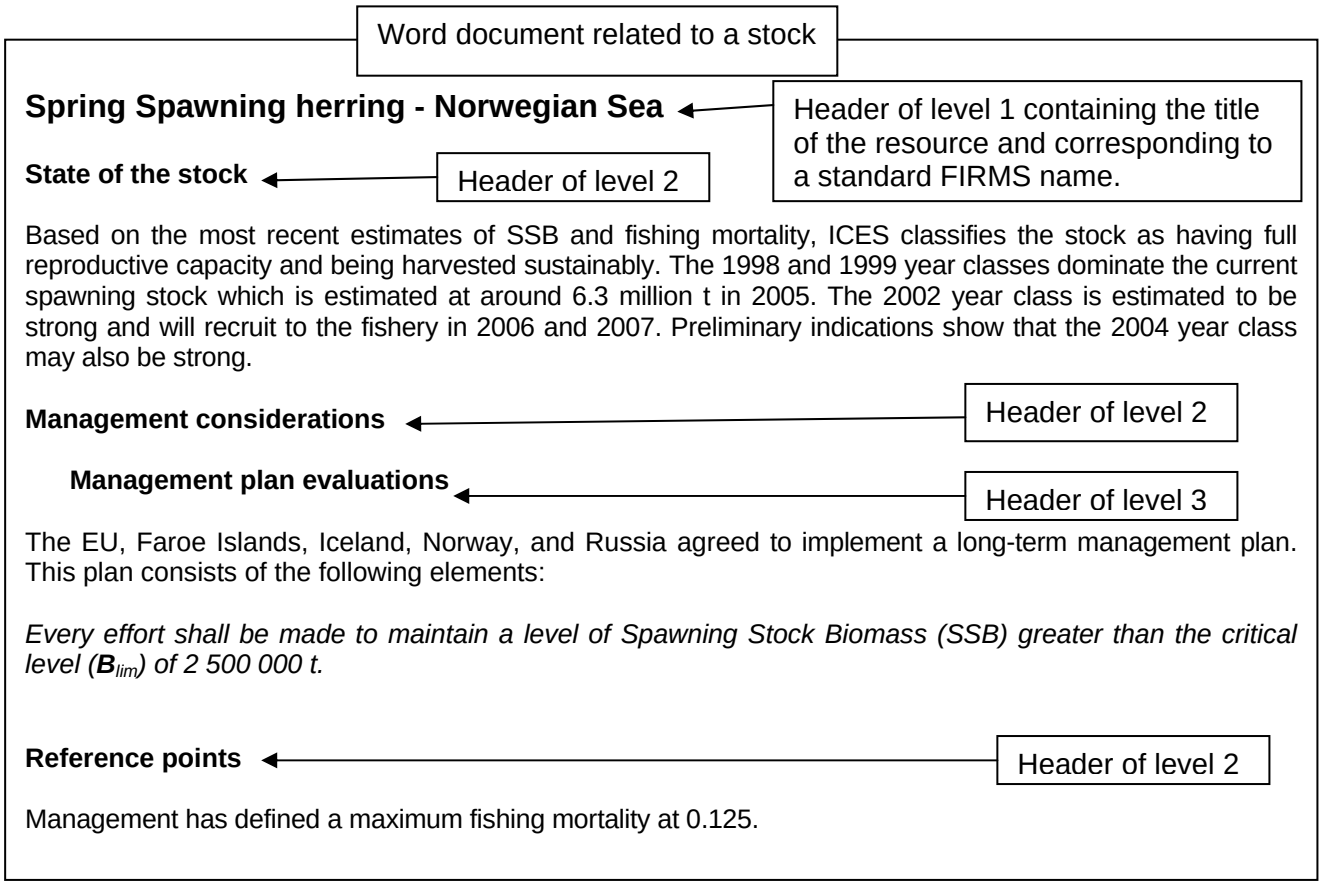
3 Conversion process: detailed description

The conversion process is held in two steps. The first step presented below was added upon request of ICES in order to facilitate quality control of the correct application of standard headers at source document level. This first step proceeds through generation of an XML file structured without semantic (semantically free). The second step therefore assumes that standard headers in the source document are 100% applied in order to prepare a FIRMS compliant XML file.

3.1 First stage: Generating a semantically free XML file from a structured Word Document

The purpose is to obtain from a structured word document an XML file with a structure that distinguishes the titles form the raw normal text, images and tables. This XML file is therefore free of semantic

Example:
This is a structured Word document provided by a partner containing the stock summary for the Norwegian spring-spawning herring resource. This document is based on a strict template with headers/titles common to all the reports created by the partner. The title of this resource “Norwegian spring-spawning herring” must correspond to a standard FIRMS name for a resource object.

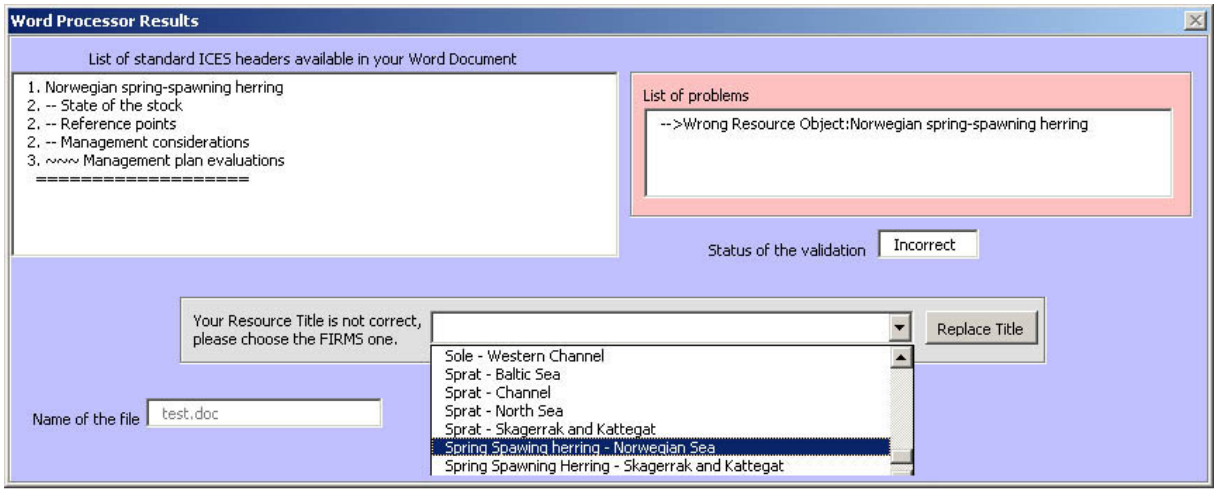


The structuring of the Word documents can be done manually or by using a tool which has been developed on purpose called “Validator”. This tool developed initially for ICES needs, generates from a non-structured flat Word document, a structured Word document with vertical hierarchy between titles corresponding to specific elements in the FiMES schema. It implies that the titles in the document match exactly the valid ones which have been listed for the Word-FiMES schema mapping.

Non-structured Word files containing relevant titles according to the agreed mapping:

- Norwegian spring-spawning herring
-
- State of the stock
- Management considerations
- Management plan evaluations
- Reference points

Using the Validator tool to automatically structure this Word document, we obtain the following interface:



It details the hierarchy which has been built based on the titles listed in the mapping between Word and FiMES elements. In the case of discrepancies such as wrong titles, the application through user-friendly views helps to modify dynamically the document with valid titles.

Result word document structured according to the pre-established mapping:

- Spring Spawning herring - Norwegian Sea
  -
    - State of the stock
    - Management considerations
      - Management plan evaluations
    - Reference points

When the document is structured with correct headers, the first conversion stage can be run. The purpose is to obtain an XML file with a structure that distinguishes the titles form the raw normal text, images and tables.

Several softwares exist to convert Word to XML. It has been decided to use upCast to generate XML files from Word reports. upCast enables to use Microsoft Word as an authoring application for XML document.

After having run the first conversion process, this is a simplified view of the XML obtained by processing the Word document with upCast. The titles of the Word document are stored into specific

tags and the structure of each paragraph is kept for the texts, images, tables and formatting initially present in the Word document.

XML file after upCast conversion obtained from an initial Word report

```
<section level="1">
  <heading class="heading 1" level="1" style="font-size: 12.0pt;"> Spring-spawning herring-Norwegian Sea </heading>
  <section level="2">
    <heading class="Normal" level="2" style="\-ilx-paragraph-outline-level: 1; font-weight: bold; font-size: 10.0pt;">State of the stock</heading>
    <par class="Normal" style="\- text-align: justify; font-size: 10.0pt; letter-spacing: -0.1pt;" xml:lang="en-US">Based on the most recent estimates of SSB and fishing mortality, ICES classifies the stock as having full reproductive capacity tnd being harvested sustainably. The 1998 and 1999 year classes dominate the current spawning stock which is estimated at around 6.3 million t in 2005. The 2002 year class is estimated to be strong and will recruit to the fishery in 2006 and 2007. Preliminary indications show that the 2004 year class may also be strong.</par>
  </section>
  <section level="2">
    <heading class="Normal" level="2" stylefont-weight: bold; font-size: 10.0pt;">Management considerations</heading>
      <heading class="Normal" level="3" stylefont-weight: bold; font-size: 10.0pt;">Management plan evaluations</heading>
    <par class="Normal" style="text-align: justify; font-size: 10.0pt;">
      The EU, Faroe Islands, Iceland, Norway, and Russia agreed to implement a long-term management plan.  This plan consists of the following elements:</par>
    <par class="Normal" style="text-align: justify; font-size: 10.0pt; font-style: italic;" xml:lang="en-US">Every effort shall be made to maintain a level of Spawning Stock Biomass (SSB) greater than the critical level (<inline style="font-weight: bold;">B</inline>
      <inline style="vertical-align: sub;">lim</inline>) of 2 500 000 t.</par>
  </section>
  <section level="2">
    <heading class="Normal" level="2" style="\; font-weight: bold; font-size: 10.0pt;">Reference points</heading>
    <par class="Normal" style="text-align: justify; font-size: 10.0pt; letter-spacing: -0.1pt;">Management has defined a maximum fishing mortality at 0.125. </par>
  </section>
</section>
```

Header of level 1

Header of level 2

Text of the header "State of the stock"

Header of level 2

Header of level 3

Text of the header "Management plan evaluations"

Header of level 2

Text of the header "Reference points"

3.2 Second stage: Generating the FIRMS compliant XML file

After the conversion of the structured document into an upCast XML file, we can generate the final FIRMS compliant XML file. It consists of an XSL transformation which allows to convert the XML generated by upCast into a new XML valid against the FiMES schema.

This transformation uses the pre-established mapping between the headers in the Word document and the elements of the FiMES schema in order to copy the content of each paragraph of the Word report into the corresponding element in the FIRMS XML.

The main title of the report (header of level 1) will be used to link this report to a specific resource object defined in the FIRMS system. This will enable assignment of the FIRMS object unique identifier.

For example, if the header "State of the stock" is found in the document, according to the mapping, the content of this header will be stored under the "fi:AqResStatusTrends" element of the FiMES schema which is the element used to describe status of the stock.

We can also make the correspondence for some others titles of the example above:

- Management considerations
  ➔ fi: BioAssessment / fi:SciAdvice

- Reference points
  ➔ fi:BioAssessment/ fi:Results/ fi:RefPoint

The goal is to fill the final XML based on a skeleton designed by the FiMES schema with the whole content of the Word document.

After transforming the upCast XML using XSL files, we get the following FIRMS compliant XML:

Final FIRMS XML obtained from an initial stock summary Word

```
<?xml version="1.0" encoding="utf-8"?>
<fi:FIGISDoc xmlns:fi="http://www.fao.org/fi/figis/devcon/" xmlns:ags="http://www.purl.org/agmes/1.1/"
xmlns:agls="http://www.naa.gov.au/recordkeeping/gov_online/agls/1.1" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.fao.org/fi/figis/devcon/ http://figis01/Dtd/Beta/3.5/firms_schema/editor/aqres_editor.xsd"
xml:lang="en">
    <fi:DataEntry>
        <fi:Editor>Bodil Chemnitz</fi:Editor>
        <dcterms:Created>2005-11-26</dcterms:Created>
    </fi:DataEntry>
    <fi:ObjectSource>
        <fi:Owner>
            <fi:CollectionRef>
                <fi:FigisID>7</fi:FigisID>
                <dc:Title>ICES Advisory Reports</dc:Title>
            </fi:CollectionRef>
        </fi:Owner>
        <fi:CorporateCoverPage>
            <fi:FigisID>7</fi:FigisID>
        </fi:CorporateCoverPage>
        <fi:CoverPage>
            <dcterms:Created>2005-01-01</dcterms:Created>
        </fi:CoverPage>
    </fi:ObjectSource>
    <fi:AqRes>
        <fi:AqResIdent>
          <dc:Title>Spring Spawning herring - Norwegian Sea</dc:Title>
            <fi:FigisID>10335</fi:FigisID>
            <fi:ReportingYear>2004</fi:ReportingYear>
                .....
        </fi:AqResIdent>
<fi:BioAssessment>
    <fi:Results>
        <fi:RefPoint>
            <fi:Text><![CDATA[<br/>Target reference points<br/><
mortality at 0.125. <br/><br/>Technical basis:<br/>]]></fi:Text>
        </fi:RefPoint>
    </fi:Results>
    <fi:SciAdvice>
        <fi:Text><![CDATA[The EU, Faroe Islands, Iceland, Norway
management plan.This plan consiwts of the following elemhall be m
(SSB) greater than the critical level (<b>B</b><sub>lim</su
    </fi: SciAdvice >
</fi:BioAssessment>
<fi:AqResStatusTrends>
            <fi:Text><![CDATA[<br/>Based on the most recent estimat
k as having full reproductive capacity tnd being harvested sustainably. T
spawning stock which is estimated at around 6.3 million t in 2005. The 2
to the fishery in 2006 and 2007. Preliminary indications show that the 2
</fi:AqResStatusTrends>
</fi:FIGISDoc>
```

Stable part of the final XML with the same structure whatever the source stock summary document is. The content of this section will be updated according to the partner (Collection and Coverpage references) and the observation (Identity block).

Name of the stock summary corresponding to a FIRMS local name allowing to extract the FIRMS identifier.

Part of the final XML that corresponds to the content of the source Word document. In our example, the Word document contains the following titles:

- State of the stock
- Management considerations
- Reference points

The corresponding XML elements containing the content of these paragraphs (text, images and tables) are present. The have been ordered to have an XML compliant with the FiMES schema.

The valid XML file and the images which the word file was containing are stored in a folder. . This constitutes a package ready for the upload in the Content Management System.

4 Overview

The resulting XML is 100% compliant with FiMES schema and respect the integrity of the content available in the initial Word documents provided by the partner.

4.1 Summary of the process

The complete process is illustrated in Annex 2.

The process to convert a Word is the following:

- To obtain a quality controlled Word document in which the headings structure is verified according to the pre-established mapping with FIRMS metadata; this can be done manually and/or by using the “Validator” tool;
- To convert Word documents using upCast into a semantically free XML file;
- Transform the resulting XML into a FIRMS compliant XML (semantically structured).

4.2 Data quality controls

A few modifications by hand using an XML editor should be done by each partner, it can differ according to the partner and the complexity of their Word document.
It is necessary to know that for the final structure of the FIRMS XML, the paragraphs of the initial Word document may be presented in a different order.

4.3 Fact sheet result

After the XML is loaded into the Content Management System, it will be viewable as a fact sheet available in the FIRMS website:

Marine Resource Fact Sheet Search Save Print

Stock status report

Spring Spawning herring - Norwegian Sea, 2004

Norwegian spring-spawning herring Citation

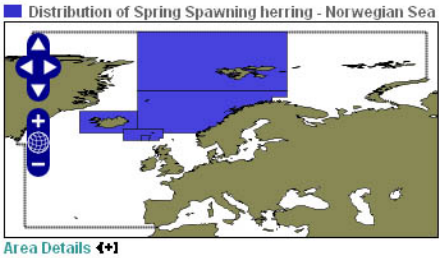
Owned by International Council for the Exploration of the Sea (ICES) more>>

Related observations Locate in inventory

Species
Clupea harengus

Fao Names : en - Atlantic herring, fr - Hareng de l'Atlantique, es - Arenque del Atlántico

Distribution of Spring Spawning herring - Norwegian Sea



Area Details

Main Descriptors
Considered a single stock: Yes **Spatial Scale:** Sub-Regional
Considered a management unit: Yes

Table of Content

Biological State and Trend

Exploitation rate: Moderate fishing mortality **Exploitation state:** Not Provided
Abundance level: Virgin or High Abundance

Spawning biomass in relation to precautionary limits	Fishing mortality in relation to precautionary limits	Fishing mortality in relation to highest yield	Comment
full reproductive capacity	Harvested sustainably	-	-

Based on the most recent estimates of SSB and fishing mortality, ICES classifies the stock as having full reproductive capacity and being harvested sustainably. The 1998 and 1999 year classes dominate the current spawning stock which is estimated at around 6.3 million t in 2005. The 2002 year class is estimated to be strong and will recruit to the fishery in 2006 and 2007. Preliminary indications show that the 2004 year class may also be strong.

Assessment

Scientific Advice

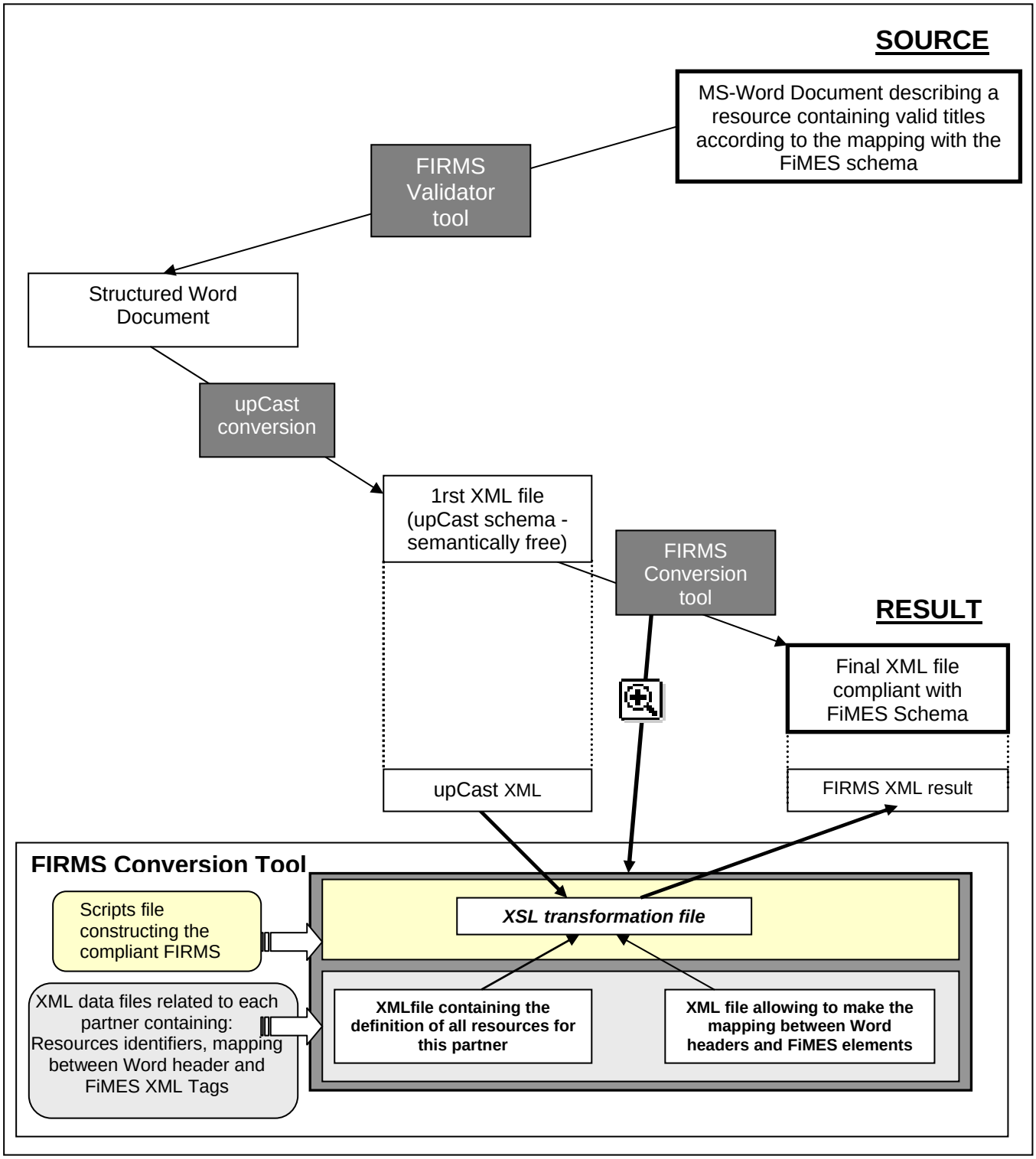
Single stock exploitation boundaries

Annex 1: Mapping

Example of mapping between ICES headers and FiMES XML elements:

ICES titles in Word documents	Header level	FiMES element		
Catch and TACs (table)	H2	fi:Assesment	fi:Results	fi:Projection
Changes in fishing technology and fishing patterns	H3	fi:Exploit		
Comparison with previous assessment and advice	H3	fi:BioAssesment	fi:Results	
Data and methods	H3	fi:BioAssesment	fi:AssesMethods	fi:AssessMethodEntry
Ecosystem considerations	H3	fi:BioAssesment	fi:SciAdvice	
Environment conditions	H3	fi:BioAssesment	fi:AssesMethods	fi:AssessMethodEntry
Exploitation boundaries in relation to existing management plans	H3	fi:BioAssesment	fi:SciAdvice	
Exploitation boundaries in relation to high long term yield, low risk of depletion of production potential and considering ecosystem effects	H3	fi:BioAssesment	fi:Results	
Exploitation boundaries in relation to precautionary limits	H3	fi:BioAssesment	fi:SciAdvice	
Factors affecting the fisheries and the stock	H2	fi:Exploit		
Fishing mortality (Standard graph)	H2	fi:BioAssesment	fi:Results	
Information from the fishing industry	H3	fi:BioAssesment	fi:AssesMethods	fi:AssessMethodEntry
Landing table	H2	fi:BioAssesment	fi:Results	
Landings (Standard graph)	H2	fi:BioAssesment	fi:Results	
Management considerations	H2	fi:BioAssesment	fi:SciAdvice	
Management objectives	H2	fi:Management	fi:Objectives	fi:ObjectiveEntry
Management plan evaluations	H3	fi:BioAssesment	fi:SciAdvice	
Other factors	H3	Exploit		
Outlook for 2006 (table)	H3	fi:BioAssesment	fi:Results	fi:Projection
Short term implications	H2	fi:BioAssesment	fi:Results	fi:Projection
Precautionary Approach reference points (table)	H2	fi:BioAssesment	fi:Results	fi:Refpoint
Recruitment (Standard graph)	H2	fi:BioAssesment	fi:Results	
Reference Points	H2	fi:BioAssesment	fi:Results	fi:Refpoint
Regulations and their effects	H3	Exploit		
Scientific basis	H2	fi:BioAssesment	fi:AssesMethods	fi:AssessMethodEntry
Single-stock exploitation boundaries	H2	fi:BioAssesment	fi:SciAdvice	
Single-stock reference points (table)	H2	fi:BioAssesment	fi:Results	fi:Refpoint
Source of information	H2	fi:Sources	fi:BiblioEntry	
Spawning Stock Biomass (Standard graph)	H2	fi:BioAssesment	fi:Results	
State of the stock	H2	fi:AqResStatusTrend		
Stock Summary	H3	fi:BioAssesment	fi:Results	
The Environment	H3	fi:Exploit		
Uncertainties in assessment and forecast	H3	fi:BioAssesment	fi:AssesMethods	fi:AssessMethodEntry
Stock-Recruitment (graph)	H2	fi:BioAssesment	fi:Results	
Precautionary approach plot	H2	fi:BioAssesment	fi:Results	
Yield and Spawning Stock Biomass per Recruit (graph)	H2	fi:BioAssesment	fi:Results	fi:Projection

Annex 2: Functional diagram



Annex 3: Glossary

XML:

XML stands for E**X**tensible **M**arkup **L**anguage.
XML is a markup language much like HTML.
XML was designed to describe data.
XML elements are not predefined. You can define your own tags.
XML was designed to describe data and to focus on what data is while HTML was designed to display data and to focus on how data looks.

XSL:

XML stands for E**X**tensible **S**tylesheet **L**anguage.
XSL is a Style Sheet Language that can be used for displaying and formatting XML documents. Using XSL is two-step process, the first step being a transformation of the XML document using XSLT, and the second step being the rendering of the result of the transformation in different formats (new structured XML, HTML…)