2nd Specific IT meeting on Task Force on Zoonoses Data Collection
- EFSA-IT Experts Meeting for the provision of Zoonoses data in XML and Excel format

MINUTES

21-22 September 2011
EFSA, Parma ITALY

AGENDA:

---

1. Welcome
2. Declaration of interests
3. Updates on Zoonoses system and XML/Excel submission:
   a. Zoonoses developments after the 1st IT Task Force
   (Fabrizio Abbinante)
   b. Feedback from the AMR sample based pilots
   (Fabrizio Abbinante)
4. Updates on DCF and submission of XML/Excel file to EFSA
      i. New DCF version
   (Jane Richardson)

5. ECDC experience on using XML/CSV file in TESSY
   (Per Rolfamre)

6. Feedback from the WG on XML/Excel use
   (Fabrizio Abbinante)
7. Suggroups
   a)   Future use of XML files for aggregated and sample based tables:
         I.   XML Schemas and XML pick lists in XML format
         II.   Maintenance of XML Schemas and XML pick lists
         III. Requirements of XML support needed and test environment
   b)
         I.   Data Model and pick lists in Excel format
         II.   Maintenance of Excel data model and Excel pick lists
         III. Requirements of Excel support needed and test environment

8. Feedback from the two subgroups and the plenary discussion
9. Conclusions and recommendations
   (Pia Makela, Fabrizio Abbinante)
10. Any Other Business

Members and other National or Commission representatives:

Arche Stéphane (FR), Bandelj Boris (SI), Hanni Andrea (CH), Colangeli Patrizia (IT), Forlizzi Luca (IT), Iannetti Simona (IT), Rydin Joao (PT), Kern Thomas (AT), Kurucz Peter (SK), Molina Carlos Alberto (ES), Nibhraonian Ruth (COMM), Stojkovska Lile (MK), O'Dea Eileen (IE), Petkov Kolayan (BG), Reinecke Annette (DE), Roosileht Reimo (EE), Sahin Guzin (TR), Saule Atis (LV), Sewell Peter (UK), Siposean Cristian (RO), Sporniak Sebastian (PL), Staskevicius Tadas (LT), Tomislav Kis (HR), Tenhagen Bernd Alois (DE), Tuominen Mikko (FI), Hedbavny Petr (CZ), Zarka Péter (HU), Per Rolfamre (ECDC)

EFSA:

Fabrizio Abbinante (Chair), Francesca Riolo, Kenneth Mulligan,Valentina Rizzi, Frank Boelaert, Simona Fusar Poli , Pia Mäkelä (BIOMO).
Stefano Cappè (DCM), Jane Richardson (SAS) Marco Leonia (ITOP).

MEETING MINUTES:

1. WELCOME

The chair welcomed the participants and announced that the intention of the meeting was to focus on future provision of Zoonoses data using XML and Excel files. The chair outlined the agenda and objectives of the meeting, introducing the various speakers, and explaining the functioning of the discussion groups.

2. DECLARATION OF INTERESTS

The chair invited four experts to fill in the Declaration of Interest form before the starting of the meeting.

3. UPDATES ON ZOONOSES SYSTEM AND XML/EXCEL SUBMISSION

3.1 Zoonoses developments after the 1st IT Task Force

Fabrizio Abbinante (FA) gave a presentation on the current provisions for the XML submission of Zoonoses data via the online web application, outlining the milestones achieved for the Food Borne Outbreaks (FBO) pilot and the subsequent extensions made to support aggregated Antimicrobial (AMR) data submission. Based on the proposals from the last IT task force meeting to:
  * Put on hold the XML development, 1 year of planning and piloting
  * launch of grants to help MSs in the implementation of XML submission
  * support of Excel submission or transformation from Excel into XML
  * planning pick list maintenances

In 2011, EFSA has set up a working group on the use of XML/EXCEL files to fully explore all technical and practical issues for the submission of data in these formats. A new and improved version of the Data Collection System (DCF) has been rolled out along with a pilot on XML and Excel submission of AMR Isolate-based data using the new DCF. Currently there are two systems in place for bulk Zoonoses data submission to EFSA, one that supports the currently adopted generic model for FBO and aggregated AMR data, and one based on a simplified flat data model for AMR Isolate-based data, which does not support complex relationships thus introducing data

redundancy that is not present in the generic model. This simple model is part of overarching data collection exercise (in DCF) which in the future will produce a unified data collection model with one harmonised pick list module. This unified model, with its harmonised pick list module, will greatly help reduce inconsistencies in control terminology, thus minimising the efforts required by member states in mapping there own pick lists to those of EFSA for all future data collection exercises.

To aid Members States who are early adopters for the submission of this type of XML and Excel data, an article 36 call to provide funds for Member States to aid the development of tools to export data to the simple DCF models has been launched. The projects that will result from this call will last 20 months. The main goal is to fund projects that will aid the collection, structuring and mapping to published EFSA data models from data collected by member states. The proposal covers all tables currently collected for the annual monitoring of Zoonoses and the AMR Isolate-based tables: the corresponding data models will be published by EFSA over the next year.

3.2 **Feedback from the AMR sample based pilot**

Fabrizio Abbinante (FA) gave feedback on the AMR sample based pilot, which was driven by the scientific need to inspect in greater details AMR data. A working group was set up, and based on their recommendations the Zoonoses Task Force supported the use of a pilot project to investigate the feasibility of data submission via XML and Excel for annual AMR Isolate-based data. FA outlined the timelines for the development phase of the data model to support this data collection, which was then distributed to the interested member states. The Data Collection Framework (DCF) which was used for the submission of collected data provided a test environment, which member states could used to test and refine their data sets for both XML and Excel. One of the main outcomes from this process was a data submission process which had few transmission error and high quality data. It was then agreed with Member States, that this submitted isolate level data could be aggregated by EFSA, and uploaded into the Annual Zoonoses collection web system. FA went into more details on the AMR data model that was developed with experts of the AMR working group, underlining the fact that is heavily based on the standard sample description (SSD) data model, which at the time of development did not support all the AMR fields needed for reporting. In the future, the current version of the SSD model will go through a proposed call for revision, extending it with the data needs of the AMR model, to allow for the collection of isolate based AMR data.

Member state Feedback

Verena Spiteller (*AGES*) gave feedback from the Excel and XML pilot, and the steps performed by the member states in the submission of data to EFSA via DCF, for both XML and Excel. Issues highlighted for XML in the pilot have been resolved. There were no major issues with data entry into the Excel file. One minor issue relates to the concatenation of both NUTS code and NUTS name. Highlighted again were mapping issues surrounding the none harmonised control terminologies between EFSA and other agencies, which leads to a greater volume of work for early adapters of the data model.

Memory issues exist in the use of Excel, whereby large files prove slow to work with and difficult to distribute. Some issues regarding the documentation of the Excel file needs to be cleared up. Naming differences exist across the different data model description documentation from XML scheme, Data Dictionary and the Excel file.

Some backward compatibility issues exist with Excel. A proposal to move correct certain xml schema message header elements/attributes from both the SSD and AMR were proposed. The use of the DCF for XML document upload and validation had no major issues, although the upload notification email was not always sent out from the production environment. Some issues

surrounding data aggregation once collected at sample level were highlighted, especially in the context of human errors within the data leading to lost data classification. The process of mapping Member State data to the models provided by EFSA was underlined as a very labour intensive, but once the mapping is concluded the resulting data extraction becomes very simple, and once done can be used in any future data collection provided to EFSA.

4.   UPDATES ON DCF AND SUBMISSION OF XML/EXCEL FILE TO EFSA

4.1.1   New DCF version

Jane Richardson (JR) outlined the current in house EFSA Data Collection Framework (DCF), explaining what the DCF does, and future improvements. In essence the DCF is a secure web portal that allows the submission of data from a wide variety of different data collections. The DCF collects both aggregated and sample based data, based upon an agreed data model. The DCF supports the collection and the storage of data, allowing easy access to the system via a unified LADP user management system. The DCF will support versioning of centralised control terminology, and will facilitate the publishing of updated and deprecated pick lists. The submission of data into the DCF is quite simple, because of the flat structures used to define the data model, whereby uploaded data is checked against the central harmonised pick lists/control terminology. The DCF in the future will support more complex business rules, where the business rules check certain heuristics that exist between variables.

Standard Sample Description data dictionary

> *(Speaker: Stefano Cappe EFSA)*

Stefano Cappe (SC) outlined the current in house Standard Sample Description (SSD), it was decided for the SSD data dictionary that excel was one of the best technologies for people where it not technically feasible for them to query a database. It has been designed in such a way that it can be easily imported into a database for those who wish implement it directly their database. The Excel file has also an Excel Forms search facility so that terms and their respective code can be retrieved. The implemented dictionaries are now currently available on the EFSA website for download, and have a stable update lifecycle of one a year. The SSD catalogs have an expert domain review panel that checks the validity of requests for new catalog entries; this process is in place to avoid duplication and validity of terminology.

5.   ECDC EXPERIENCE ON USING XML/CSV FILE IN TESSY

(Speaker: Per Rolfhamre ECDC)

Per Rolfhamre (*ECDC*) gave a presentation on experience gained through the use of the Tessy system, the mission statement behind which is to strengthen European surveillance of human infectious diseases

The introduction of the Tessy system coincided with a nominations process of a national coordinator as national contact points. Each disease has a surveillance network. Tessy is a system for the collection of European surveillance data, in effect a one stop shop for all European stakeholders. The millstones of the Tessy system were presented with its inception date starting in 2006. There is also a training program that is in place to teach operators on how to upload data collections. Every year Tessy supports ninety data collections, and to date has more than twenty three million records which are a mixture of case based and aggregated data. Some of the data

collected is reported online, but is mainly available to domain specific users in the system. An overview of the different files types that are currently collected and validated in the system was given. The clean and validated data in the system is only inserted into the database when users are accepting of the cleaned data produced by the Tessy system. The use of XML is very low, making up only 9% annual data, with the predominate format used being Comma Separated Values (CSV). Manual data entry is used significantly when the data collections are low in data frequency. The use of XML is limited to only eight countries, with again CSV being supported by 50 countries. The lessons learned especially related for sample data, surrounds the problem where unique identifiers, simply put reporting countries cannot support unique identifiers (IDs) and as such Create , Read, Update, Delete (CRUD) functionality at individual level items is not possible. The CRUD functionality for sample based data is supported by the use of time based reporting periods, and aggregated data is only updated via a complete replace of data in the system.

6. FEEDBACK FROM THE WG ON XML/EXCEL USE

Fabrizio Abbinante gave a detailed run through of the working group set up to monitor the pilot for the AMR and to check the outcomes of the pilot considering how to use XML and Excel for the submission of this data. One of the main tasks of the working groups is the maintenance and technical support to be provided to Members States for both the XML and Excel data collections and submission vehicles. In all, the WG will have five meetings by the end of this year, producing a technical report that will be presented to the Zoonoses Task Force. Some of the current outcomes focus on:

- The need to extend the SSD to include missing filed needed in the collection of AMR at isolate level.

- The future use of DCF for the submission of both AMR and Zoonoses data in Excel and XML format.

- To use the simple flat data model, and to drop the more complex generic data model.

- To merge similar Zoonoses tables into the same data model like for Serovar and Phagetypes tables.

- Support up-to-date distribution of amended pick lists.

- To provide training to member states to help with both mapping and submission of data collections.

7. DIVISION IN THREE SUBGROUPS FOR DISCUSSING THE FOLLOWING TOPICS

7.1 Future use of XML files for aggregated and sample based tables

The chair invited the all present to create two subgroups composed of 10-15 experts who where to discuss the following topics:

7.1.1 XML Schemas and XML pick lists in XML format

7.1.2 Maintenance of XML Schemas and XML pick lists

7.1.3 Requirements of XML support needed and test environment

7.2 Future use of Excel files for aggregated and sample based tables

7.2.1 XML Schemas and XML pick lists in Excel format

7.2.2 Maintenance of Excel data model and Excel pick lists

7.2.3 Requirements of Excel support needed and test environment

8. FEEDBACK FROM THE TWO SUBGROUPS

The two Rapporteurs presented the outcomes of the discussions to the entire audience.

The subgroup on "Future use of XML files for aggregated and sample based tables" agreed the following proposals

Generic vs. Simple data models

- MS argue strongly in favour for the flat model, so this should be preferred.
- That means there will be different models for the different topics

Combining picklists

- Sometimes combination of columns is useful, (eg zoonosis)
- Others are combinations of different concepts (sampling strategy) that may have to be handled separately

Maintenance of pick lists

- Sometimes terms have to be included in pick lists during the reporting season
- No issue of other MS but requests may occur in duplicate – something EFSA must handle
- A lot of things can be included in the pick lists provided at the beginning of the reporting season, but specific things can occur later.
- Web service was propsed to inform MSs on changes in a requested time period.

Data Transmission

- will be organized according to the published guidelines (already agreed by TF)
- Access to the test environment for more people
- Pick list outside of schema because schema would have to be changed to often.
- Testing against pick list should be available in the testing environment.

Discussion on proposals

Fabrizio Abbinante (FA) explained that from a technical prospective there is no problem in providing pick lists in various formats. There is a need to revisit some of the pick lists regarding, for example, sampling context, where perhaps a combination or redesign is suggested to facilitate the mapping of member states sampling context information in more detail.

The proposed data submission by EFSA via the DCF was accepted.

The subgroup on "Future use of Excel files for aggregated and sample based tables" agreed the following proposals

Excel

- The group supports the submission of Zoonoses data using Excel format
- Excel is the major data management tool for many reporting organisations
- It allow reporting officers to correct mistakes when a transmission fails validation
- Macros for generating XML can present a security problem

Templates

- The group would like EFSA to prepare Excel templates to support reporting of Zoonoses data
- Reporting organisation to maintain mapping from national terms to pick lists
- To reduce the number of templates there could be a review of the format of the tables with consolidation and consideration of national organisation report formats

IT Task Force

- To approve templates for each reporting season
- To maintain pick lists and discuss unmatched terms in national systems
- Frequency – annual

Other issues

- For aggregated data file size is not an issue
- Test environment essential
- Mapping training in excel – Online training material available from ECDC

The conclusions of the two subgroups were supported by the entire audience with no modifications.

9. CONCLUSIONS AND RECOMMENDATIONS

All Member States were invited to send offers for next years article 36 grants call, in order to receive support for the automatic transfer of data..

In order to help the MSs in getting familiar with the simple data models, FA suggested the possibility to send upon request a "country specific" historical flat file with all the data submitted by the requesting country in the Zoonoses database.

The chair announced that from EFSA's part the meeting was really useful and successful. A report from this meeting will be given by Fabrizio Abbinante at the next Zoonoses Task Force meeting on 28 October 2011. The following proposals to the Zoonoses Task Force will be made for approval:

- Use of "simple flat models" for both aggregated and sample based data
- Use of DCF for future XML/Excel/CSV files submissions to Zoonoses
- Automatic migration of data from DCF database into the Zoonoses Database
- Automatic aggregation and migration of isolate-based data from DCF into Zoonoses
- For DCF, Zoonoses and all future data collection projects in EFSA, shared and common:
    - exchange protocol
    - dictionary management module
    - user management module
- Support in the mapping of national pick lists against EFSA's pick lists but not implementation of Excel Macro from EFSA: training, coding guidelines and suggestions
- Update mechanism (not only annual but also during the reporting season) to refresh the published pick lists and notification mechanism for informing the MSs about updated pick lists
- Suggestion to revise of the Zoonoses pick list that contains "several" scientific concept inside the same pick list hierarchy (e.g. Sampling Context, Sampling Stage)

The suggestions from today's meeting will be used in the development of the Zoonoses data models, where the models and pick lists will be published as early as possible. Access to the DCF and all required tables as well as a test environment will also be provided so that member states start as soon as possible.

10. ANY OTHER BUSINESS

None