European Food Safety Authority

EVIDENCE MANAGEMENT UNIT

DIGITAL TRANSFORMATION SERVICES UNIT

ASSESSMENT AND METHODOLOGICAL SUPPORT UNIT

Concept paper on the future of scientific data in EFSA

1. Background

Data are at the heart of EFSA's 2020 Strategy. Within the framework of strategic objective 2, "Widen EFSA's evidence base and maximise access to its data", several cross unit activities are ongoing to deliver three underpinning operational objectives[1]: adopt an open data approach, ii) improve data interoperability to facilitate data exchange and, iii) migrate towards structured scientific data.

EFSA has already started to make much of its data and evidence publicly accessible via its scientific data warehouse, knowledge junction, repositioned EFSA journal on Wiley, as well as European open data portals. EFSA's published outputs are now available as JATS[2] XML, the international standard for journal articles. Within the context of the MATRIX project, EFSA is in the process of piloting migration from PDF dossier applications in the regulated product area towards electronic dossier submission and automatic publication of non-confidential information using structured formats based, insofar as possible, on existing international standards to enable data access and re-use.

This paper looks beyond the current EFSA strategy and its implementation plan. It reflects on the acquisition, management and use of data in EFSA by 2025 and beyond, considering developments in information technology and digitalisation in the environment in which the agency operates. This includes recent developments like machine learning and collective intelligence available globally via the internet that are capable of searching, elaborating and digesting enormous quantities of data.

In particular, this paper elaborates on several initiatives that were considered of significant potential and feasibility following two workshops on this topic in Q2 2017 with staff from AMU, DATA and DTS. An overview of shortlisted initiatives clustered in four thematic areas is presented together with examples of scientific needs that they could help to address, as well as key enablers to do so.

It is envisaged that the content of this paper will contribute to future strategic planning concerning data at EFSA.

2. Thematic areas

2.1 Scientific innovation and new data streams

It is widely acknowledged that some 90% of the data in general in the world today has been created in the last two years and about 75% of these data is unstructured.

EFSA is already active in curation and dissemination of toxicological data used in its scientific assessments through the OpenFoodTox dashboard. Novel information streams using crowdsourcing and associated IT platforms are being explored by EFSA to complement existing

[1] https://www.efsa.europa.eu/en/corporate/pub/amp1719

[2] Journal Article Tag Suite

information sources used to produce scientific assessments. In addition, the first phase of the transformational MATRIX project is reaching completion. Under the MATRIX project EFSA is seeking to combine different XML[3] standards (GHSTS, OHTs and SSD[4]) within the regulatory dossier package. This will allow stakeholders to exchange information at different levels of detail from simple product summary information to individual blood test and weight measurements from rats included in repeated dose toxicity studies.

Many datasets have also become so large and complex that we need new tools and approaches to make the most of them. Next generation sequencing (NGS) is revolutionising our understanding of biological systems and tools such as CRISPR[5] will allow us to use this information in powerful ways. EFSA is already using lessons learned from the molecular typing project in tackling a new Commission mandate on whole genome sequencing. Whole crop plant sequencing data easily represent the largest files that EFSA has the legal obligation to handle and store. When the Matrix project is fully operational, this will create a growing pool of scientific data which will require a robust long term solution to store and integrate with existing data sources. To ensure best use of these data it is essential that the information is stored in systems which allow users a single point of access to the data with tools for visualisation, annotation and quantitative analysis.

Indeed, many new in vitro methods are in development to improve our understanding of the effects of chemicals and pathogens at the cellular and molecular level. While EFSA's newly established knowledge and innovation community (KIC) on *biotechnology and molecular methodology* aims to integrate these methods in a unified way across units, EFSA needs to be able to access and process these kinds of large, complex datasets with minimal manual interventions.

It is not just laboratories that are generating big data[6]; throughout the food chain there is an increasing number of real-time monitoring systems from smart tractors which can measure and control the volume of pesticides applied to crops to automatic scales in beehives and milking parlours which increases exponentially due to the 'Internet of things' where every sensor, machinery and gadget becomes an on line data generator. This information would be extremely valuable for EFSA to gain access to define scenarios, refine risk assessments or measure the impact of emerging risks or new control methods.

Exploration of all plausible data streams, including from the general public, could generate useful information to inform our future scientific work ~~.~~In order to be able to source knowledge and information available in the public domain existing IT platforms need to be explored or developed in order to facilitate the harvesting and exchange of data and ideas using crowdsourcing. An assessment of data quality (fitness for purpose) prior to use is a prerequisite in this regard.

Enablers:

- *OECD chemical safety products:* Global Harmonised Submission Transport Standard, Harmonised templates for reporting chemical test summaries, QSAR toolbox and MetaPath[7] platform.
- *EFSA scientific data warehouse*.
- Open source analytics platforms such as KNIME
- *EFSA's DEMETER project*: establishment of an EU emerging risks platform to strengthen EU risk assessment capability.
- *Current collaboration activities with ECHA and EMA:* related to "increased sharing and publication of scientific non-clinical safety data in a harmonised format".
- *European Molecular Biology Laboratory (EMBL) databases and bioinformatics toolboxes.*
- *Access to real-time monitoring systems within the food chain (e.g. the Internet of Things)*.
- *European Commission Digital Single Market & ISA^2 programme.*[8]

[3] XML: Extensible Markup Language - a format to store and exchange data.
[4] GHSTS: OECD Global Harmonised Submission and Transport Standard; OHTs: OCEC harmonised templates; SSD: EFSA Standard Sample Description data model. GHSTS and OHTs are standard formats for reporting information used for the risk assessment of chemicals.
[5] CRISPR: clustered regularly interspaced short palindromic repeats.
[6] In the context of this concept paper, *big data* refers to data that don't fit comfortably – or at all – into EFSA's current data management system.
[7] MetaPath platform: collects, organises and analyses experimental data from metabolism studies.

- *Member State stakeholder engagement approaches such as the [EU bee partnership] to improve the collection and sharing of data on indicators of bee health.*
- National institutes dealing with (big) data, e.g. Oxford Big Data Institute, the Alan Turing Institute, the Centre for Agricultural Informatics and Sustainability Metrics Centre for Agricultural Informatics and Metrics of Sustainability.

2.2 Distributed data: from 'data collection' to 'data connection'

Increasingly, the nature of EFSA's scientific work requires access to data not traditionally collected by the agency. As more data (and the technology to access them) are increasingly available, it is timely to consider a shift in focus from 'data collection' to 'data connection'. An API (Application Processing Interface) is the back-end technology to facilitate this transition.

Whereas the EFSA website represents the main 'shop front' to EFSA's data and information for stakeholders, an API represents an electronic 'shop front' to EFSA's data for machines. APIs are becoming more mainstream in everyday consumer transactions (e.g. hotel bookings via booking.com), and are increasingly being used by organisations to allow automatic exchange of information via the internet without the need for human intervention. In EFSA's context, APIs are an essential digital tool to improve accessibility to EFSA's data and to automatically connect to and retrieve data sources from outside the agency.

The Wiley platform, which hosts the EFSA Journal, and the Zenodo platform, which hosts the EFSA Knowledge Junction community, both have open access APIs that allow automatic information exchange between portals and other applications. As a consequence, EFSA has benefitted from increased access to its outputs and supporting material via these third party APIs. To date, however, EFSA has not yet implemented its own API to automatically 'expose' its onsite data.

Ultimately, an ecosystem (virtual network) of APIs has the potential to provide EFSA and its stakeholders with access to up to date, relevant data without duplication and storage overheads. Each data creator in the ecosystem would collect, validate, store, maintain and operate appropriate access controls. It is plausible to imagine an API ecosystem where requests and responses are exchanged between stakeholders for packets of data to parameterise models, add layers to maps or even to parameterise a request for information from a different service. Consider the case of anti-microbial resistance (AMR) for example: EMA authorises the use of antibiotics, sets MRLs (maximum residue limits) and monitors for adverse effects; food laboratories are testing for AMR substances in animals at farm and slaughter and for micro-organisms in food with resistance to antibiotics; medical laboratories are testing for antimicrobial resistance in bacterial samples from patients; health authorities monitor prescriptions of antibiotics, veterinary authorities track sales of antibiotics and patients and farmers can report treatment failures. If all stakeholders were able to rapidly and frequently exchange this information and combine this with data driven methodologies it may be possible to elucidate the factors which result in increased AMR occurring and, more importantly, to monitor the effectiveness of control strategies. In the regulated product area, it is plausible to envisage that when evaluating regulated product dossiers underpinned by large datasets (e.g. NGS) EFSA would access the data from securely maintained third party data repositories.

EFSA needs to develop and implement its own APIs to bring openness and connectivity to the next level. In doing so, the agency would establish itself as a forerunner public administration in this area.

Enablers:

- *An EFSA API*: *an electronic shopfront to EFSA's data for machines.*
- *EU open data strategy*: embracement of open data policies among stakeholders.
- *Adoption of interoperability standards*.
- *A data scientists' community of knowledge:* A community of data managers and scientists to support this approach currently does not exist but it is in the interest of EFSA and national competent authorities that the competencies around the food safety data are

[8] ISA^2: Interoperability solutions for public administrations, https://ec.europa.eu/isa2/home_en

clustered and better exploited. Cross fertilisation from other areas such as industries and universities could improve the competency area. Services of digital collaboration, code sharing, webinars and web meetings could facilitate the setup of such a community. The adoption of data standards and data quality standards are important in this regard.

- Partnerships with Member State competent authorities and EU agencies.
- Co-location of IT, data and science staff to drive forward.

2.3 Quantitative and data driven methods

Transforming (big) data into scientific evidence

In EFSA's context, a piece of evidence for an assessment is represented by data that are deemed relevant for its specific objectives. The journey from data to evidence follows an itinerary starting from selecting and collecting to appraising and validating, and finally analysing and integrating[9].

Most data driven methods require data processing pipelines to clean and transform the data for analysis. The true power of data driven methodologies is only obtained when data from multiple sources are combined and this requires interoperability standards and domain ontologies. To add to this, artificial intelligence extends the power of computers to tasks that were originally performed by humans. Artificial intelligence such as machine learning and natural language processing can discover patterns and relationships in information from millions of texts, books, online articles and other sources (e.g. social media), harvesting information that could take researchers (humans) decades to discover, retrieve and digest.

EFSA has already started trying to integrate some advances in these areas in organisational and scientific tasks; it has started a KIC on *automation, machine learning and artificial intelligence* in order to foster innovation and co-operation between EFSA staff, experts and stakeholders and to promote collaboration on the implementation of machine learning with sister agencies.

Big data offers the opportunity to test more complex predictive models and machine learning techniques (MLT) in the field of risk assessment. MLT are well suited to multifactorial problems and have been applied to animal based measure datasets to better understand tail biting in pigs and welfare in dairy cattle. R4EU has also implemented Bayesian network analysis and conditional classification as part of a set of tools to analyse multiple drug resistance patterns. EFSA's DEMETER project (Determination and metrics of emerging risks) is also implementing Bayesian network analysis to integrate expert knowledge elicitation with citizen driven emerging risk identification to infer potential emerging risks. EFSA's current collaboration project with BfR ensures access for competent authorities to *Food Chain Lab* which uses a network model to trace food contamination events. In addition, work is ongoing on the application of MLT to the screening and classification of literature to support systematic literature reviews.

Enablers:

- *Machine learning techniques*.
  Unstructured data mining: much of EFSA's evidence is unstructured (e.g. scientific articles in PDF format) and not easily accessible. EFSA already uses unstructured data mining in several projects, e.g. the MEDISYS media monitoring system provides EFSA tailored event-based surveillance in the area of plant health, and the TNO ERIS (Emerging Risk Identification Support System) text mining tool has been tested to support emerging risk identification for specific fish species. Application of a new technology, similar to that underpinning API development, allows us to envisage a process where structured information is automatically extracted from unstructured data and subsequently inserted in a database available to EFSA staff immediately for searching and querying.
- *A Standardised Scalable Collaborative Data Management Cloud Solution*: As the amount of available scientific data is growing larger and larger making impractical, for a single organization, to manage, analyse and share their datasets, a possible solution could be the implementation of a "Science-as-a-Service" platform, where one or more organizations use shared cloud resources and tools creating an interoperable community, e.g. Member states could directly prepare the data for control and monitoring in the same environment of

[9] EFSA Prometheus initiative: https://www.efsa.europa.eu/interactive_pages/prometheus/prometheus

EFSA, as well as sister agencies sharing the same architectural needs. A common cloud solution with sister agencies should be explored in this regard.

- *Multidisciplinary teams* of mathematicians, biometricians, bioinformaticians, computational toxicologists, predictive modellers, data scientists and risk assessors with access to big data would allow EFSA to tackle complex problems such as the landscape level impact of multiple stressors on ecosystem services.

2.4 Exploring the living opinion

From static PDF documents to real time analysis and communication

EFSA's stakeholders have called for its work to be more clearly communicated[10]. With new technologies entering the mainstream, disruption to traditional ways of producing outputs is inevitable.

EFSA has already started moving in this way. The recent publication on EFSA's website of 36 interactive Storymaps[11] in conjunction with the scientific opinion on vector borne diseases is a good example of the future direction for EFSA outputs. In this project information from scientific literature was systematically extracted and presented in Microstrategy reports linking data to source publication with filtering and sorting functionalities. Structured data collected from the joint EFSA/ECDC Vectornet project and from disease surveillance and alert systems was combined and presented in ArcGIS maps. The data sources were combined with expert opinion and using the MintRisk[12] open source model the risk of introduction for 36 vector borne diseases was assessed semi-quantitatively. Users can explore all components to the risk assessment through the Storymaps and they can be easily updated as and when new information becomes available. With the increase in availability of structured data and the implementation of data driven methods the next step is to make EFSA risk assessments available in an interactive format with a focus on reproducibility and transparency. These considerations will be a key component of the planned Linked EFSA Journal project currently in an envisioning phase.

It is envisaged that the scientific data warehouse would be the cornerstone for data visualisation services. The goal would be to combine standardised data from surveillance, monitoring and control activities or standardised parts of the application dossiers with ad hoc data collections undertaken to support specific scientific issues and mandates, by applying a standard approach to describing time, geography, hazards and food/animals/crops. This would underpin dashboards for standardised indicators as compliance of samples and analytical results, residues definitions, food and feed legal classes, as well as toxicological end points. Integration with the R4EU statistical toolbox would support geospatial analyses and predefined exposure scenario analysis. Public access will be through the EFSA Journal with a focus on reducing text and increasing descriptive interactive visualisations.

We can imagine that the viewer selects a substance and is shown a timeline of risk assessments with the ADIs (Acceptable Daily Intakes) set by the panels. By clicking on an ADI the viewer is presented with the evidence used in the risk assessment, with the result of the working group appraisal of the evidence and a link to all publically available sources. Selecting the sources used to set the reference point the viewer can access the R4EU bench mark dose tool and view the results used in the assessment. Alternatively selecting the sources reporting concentrations/uses in food the viewer can explore the different levels of the substance in different food groups. Finally, by selecting the sources used in setting the reference value the viewer can access the consumption data used in the assessment and is shown graphically the exposure estimates (and uncertainty around those estimates) considering different scenarios.

Enablers:

- Knowledge Junction: R4EU the open access statistical tool box.
- Scientific data warehouse.
- Linked EFSA Journal project.

[10] http://www.efsa.europa.eu/sites/default/files/efsa_rep/blobserver_assets/shpdgtiSummaryReport2014.pdf

[11] https://www.efsa.europa.eu/en/press/news/170511

[12] Mintrisk: Method for INTegrated RISK assessment for infectious diseases in animals.

- Semantic web technologies (RDF, OWL, SKOS SPARQL etc.), JSON-Linked Data and graphical databases.

3. Conclusion and next steps

This concept paper describes four thematic areas relating to scientific data where we believe EFSA should develop, applying developments in information technology, in order to remain agile, relevant and connected in 2025 and beyond.

The paper is pitched at a high level. It is envisaged that the areas described would be further refined considering unit needs mapped to each thematic area (Annex I) to constitute key elements of EFSA's 2025 strategy, and shape the future direction of scientific assessments undertaken at EFSA.

Document history	
Document reference	1.2 Final draft incorporating comments received following an internal consultation of all scientific units on the draft paper (July – Sept. 2017)
Prepared by	J. Richardson, A. Scotto, S. Cappe, D. Verloo & M. Gilsenan (with input from A. Healy)
Reviewed by	M. Gilsenan
Last date modified	22 December 2017

ANNEX I: Unit needs mapped according to thematic areas[13]

	Thematic area 1 **Scientific innovation and new data streams**	**Thematic area 2** **Distributed data: from data collection to data connection**	**Thematic area 3** **Quantitative and data driven methods**	**Thematic area 4** **Exploring the living opinion**
FEED		Data scientists' community of knowledge: Need to develop a community including MSs, academia, companies and producers. Purpose: continuous exchange of data in particular in the FEED area where the current guidance documents lack in some cases suitable technical updates. The platform would also serve as a regular basis for sharing documents	Shared cloud resources and tools creating an interoperable community: Strictly linked to the community of knowledge could be this point in the thematic area 3. Though in FEED there is a limited number of "official" data available (dossiers and not that data rich) there is a lot of info available in different contexts (researchers, companies…) which we should try to get	

[13] Needs were added by units during a consultation of a draft of this paper (July –Sept. 2017)

		(e.g. sort of continuous public consultation/peer review)	access to and to analyse with the ultimate goal to increase the quality of the risk assessment.	
SCER	Integrate different data streams for identification of emerging issues Bayesian network analysis for modelling of complex systemic risks assessment Natural language processing and machine learning for improved web screening and validation of potential emerging issues	Establishment of a EU emerging risks exchange platform (ERKEP)		RASFF model for risk evaluation of food and feed contaminants
NUTRI		Collection of consumption data for infants 0-3 months old.		EFSA has published 32 opinions on dietary reference values and two more are to be produced. It is difficult for readers to get an overview of the work and to navigate through the different opinions. Currently, the Unit is preparing a summary report which includes summary tables and hyperlinks. It would be useful if the opinions and the summary report would be presented in a more living format. It would be worth to test the publication of highly sensitive opinions in a more interactive format.

GMO	Transforming existing databases (e.g. link) into a dynamic form to facilitate quick extraction and analysis of relevant information on pests for GM plants, e.g. crop specific	Collection of data on the levels of newly expressed proteins (NEP) with integrated tools to perform meta-analysis	DNA sequencing data (genomics) are currently being submitted in applications. In order to analyze and interpret this vast amounts of raw data produced from such methodology structured processing pipelines would be needed. Data from other –omics approaches (e.g. transcriptomics, proteomics) might also become part of the risk assessment in the near future. Similar structured processing pipelines might be needed as with genomics. Data from several –omics approaches can be combined to realize their great potential.	
		Collection of toxicological data (28-day repeated dose and 90-day feeding study)	Data from multidisciplinary approaches (e.g. –omics, activity assays, etc.) can be integrated and combined with modern in-silico methodologies such as ”Machine learning” for the analysis, for instance, of biological pathways, prediction of protein structural/ functional characteristics, etc.	
		Agronomic, phenotypic, landscape, and compositional data of different crops subject of the applications for authorization (e.g. micro and macronutrients, phytochemicals, etc.)		
FIP	Could involve powerful QSAR toolboxes that could help us in establishing structural-activity		Extraction of quantitative information from the food labels available in the Mintel database	

	relationships in our evaluations and PBTK models that could help us to estimate the inner exposure to flavourings			

Glossary of IT terms[14]

Artificial Intelligence (AI) is an umbrella term covering a broad field of study with the following major subfields: machine learning, neural networks, deep learning, cognitive computing and natural language processing.

AI works by combining large amounts of data with fast, iterative processing and intelligent algorithms, allowing software to learn automatically from patterns or features in the data. AI makes it possible for machines to learn from experience, adjust to new inputs and perform human-like tasks. It has become popular today due to increased data volumes, advanced algorithms, and improvements in computing power and storage.

Machine learning, a subfield of AI, is a technology that allows computers to perform specific tasks intelligently, by learning from examples. It enables researchers, data scientists, engineers and analysts to construct algorithms that can learn from and make predictions based on data. Rather than following a specific set of rules or instructions, an algorithm is trained to spot patterns in large amounts of data. Data are therefore the fuel from which machines make predictions. Common applications of machine learning include driverless cars, credit card fraud software and voice recognition software.

Deep learning is a machine learning method that takes this idea further, by processing information in layers where the result/output from one layer becomes the input for the next one. Common applications include image and speech recognition.

Natural language processing (NLP) is a form of machine learning that allows computers to process written or verbal information. The next stage of NLP is natural language *interaction*, which allows humans to communicate with computers using normal, everyday language to perform tasks.

Cognitive analytics is the application of the above technologies to make decisions.

[14] References: The Royal Society (2017), Machine learning: the power and promise of computers that learn by example; IBM (2017), Machine learning for dummies; Deloitte (2014), Tech Trends 2014: Cognitive Analytics; SAS, Artificial intelligence, machine learning, deep learning and beyond, accessed Nov. 2017.

.