Making Best Use of Structured Data

73rd meeting of the EFSA Advisory Forum

Luisa Consolini
Information Systems Director –
European Chemicals Agency (ECHA)

ECHA EUROPEAN CHEMICALS AGENCY

Chemicals Legislations Operated by ECHA

REACH
Registration
Evaluation
Authorisation

All chemicals
>1 tonne
per annum

CLP
Classification
Labelling
Packaging

All chemicals
and mixtures

UN-wide
standards

BPR
Biocidal Product
Regulation

Active substances
and biocidal
products

PIC
Prior Informed
Consent

Import/export of
certain hazardous
chemicals

Rotterdam
Convention

Main source of information

Companies collect or generate information on properties and uses of their chemicals, assess the risks and recommend safety measures

Such information is submitted according to the information requirements set in the regulations, **IN IUCLID FORMAT**

ECHA
EUROPEAN CHEMICALS AGENCY

IUCLID FORMATS?

No Tool, No Standard

- IUCLID is an IT tool for the management of chemical safety data using the OECD harmonised templates as its data standard
- Companies use it for REACH, CLP (including Poison centre notification under Art.45) and BPR
- The submission of the information requirements to ECHA known as "Waste database" will be based on IUCLID
- ECHA promotes IUCLID towards international regulatory agencies (e.g. Australia, New Zealand, Canada)

IUCLID 6

IUCLID is the main vehicle for structuring

chemical safety data

For example:

Substance identification

The essential bit of full toxicity studies (robust study summaries)

What Does It Mean?

- A data field valued "NO Effect Level" in IUCLID

Is distilled by the submitter of a IUCLID dossier

from a toxicity study

in which tests, exposure and expert judgement are described in a non structured manner (e.g. free text)

In some cases full text documents are handled as attachments, i.e. non structured information

A refinement of the information requirements in the regulations can generate an update of the IUCLID formats and therefore more structure into previously absent or non structured information:

e.g. substances in nano form

Data (format) harmonisation

- Standardisation of data requirements by agreeing on a common definition and format for data elements
- Definition of an electronic format following the standard (XML)
- Examples in a chemical risk assessment context:
  - OECD Harmonised Templates
  - GHSTS

Some Figures

“First Collection Era”

- REACH Registration: 22 000 substances covered
- CLP Notifications: 148 000 substances

All data in ECHA is in the IUCLID/OECD Harmonised Templates

384.4 GB (database); 162.5 k Files; 300 M Fields; 5.4 M Attachments

Plus less “IUCLID friendly” data

BPR active substance data

Other sources

Legacy data

What Do We Do With it? Efficiency

- Apply validation rules
- Automate the calculation of fees
- Automate the dissemination of public information in the Chemicals Info Cards and Brief Profiles
- Simulate the above before submission
- Generate Reports
- Automate the processing to issues decisions (e.g. registration decisions)
- Searches

What Do We Do With it? Augment

- ECHA has been supporting the development of the OECD QSAR Toolbox: a toolbox to predict toxicity, including grouping and similarity functionalities
- ECHA has developed CHESAR to support Chemical Safety Assessment
- *Others could develop new capabilities based on IUCLID data*

What Do We Do With it?

Data Analytics

- More advanced searches
- Machine based screening for prioritising cases for regulatory work
- Grouping of substances

Example of capabilities

- Algorithms for structural similarity
- Hierarchical clustering of substance X and derivatives
- Analysis of structural diversity into a cluster
- Advanced visualisation on high data volumes

What Can We Do With it?

Interoperability

- Joint regulatory work with other international agencies
- Import of semi-structured data
- Data value discovery
- Export of datasets of public nature
- …

Some Figures

"Second Collection Era". It Is Also About Chemicals!

We have started the second big collection of structured data (including reverse engineering):

~15000 case data generated by doing regulatory assessments (e.g. dossier compliance checks, testing proposal evaluation, substance evaluation, restrictions, authorisations, identification of SVHC etc.)

The legacy data was "mildly structured" and we missed information

e.g. *more structure and more granularity*

Assessment outcome recording used to be at case level

It is now recorded for each hazard and use finding within a case

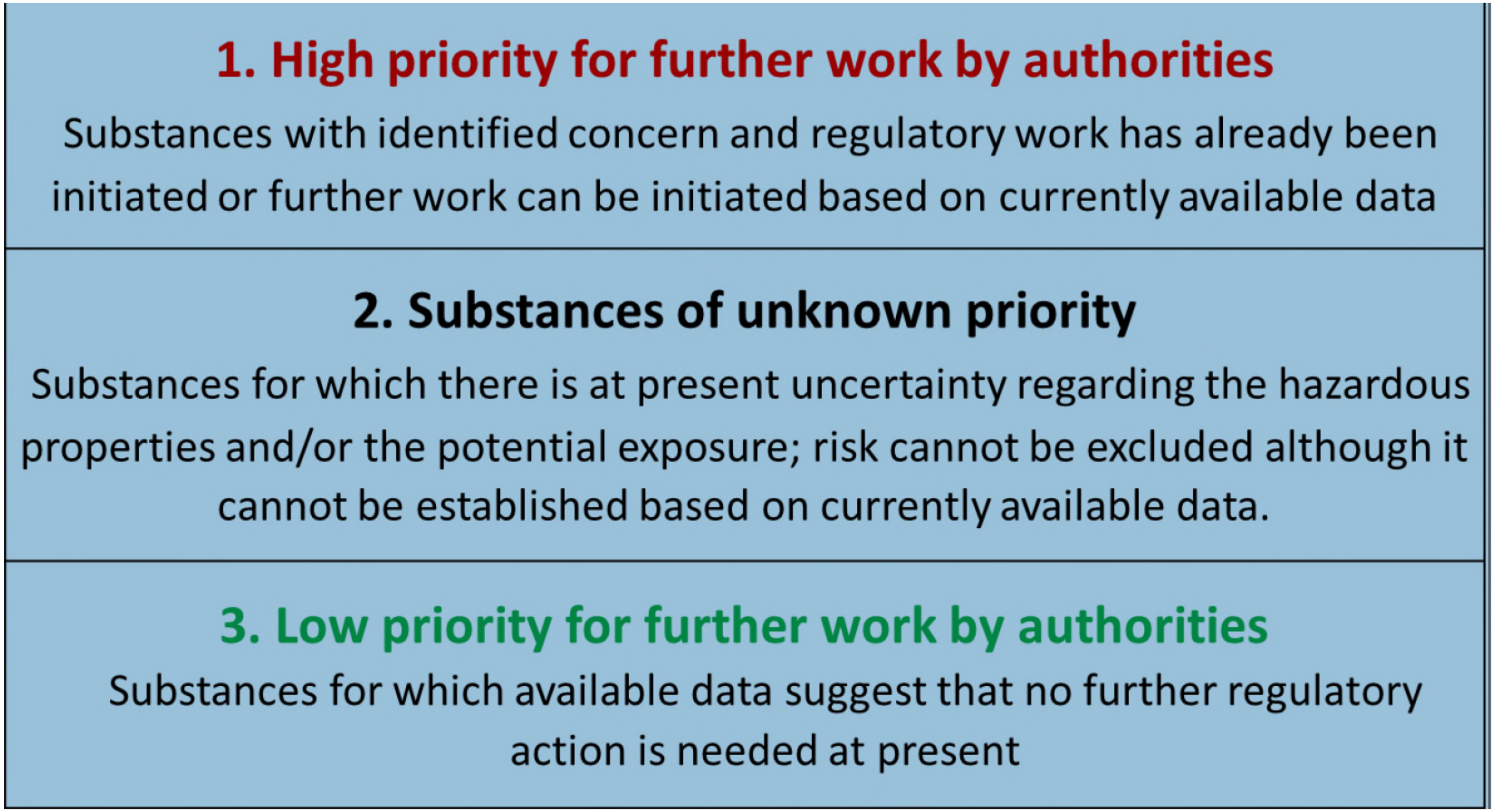
What Do We Do With it?

Consistency, Traceability, Integration… explicit knowledge

- Follow one hazard of interest throughout the REACH (and CLH) processes.
  - Monitor how a regulatory concern has evolved over time based on the information recorded by individual processes
- Record more harmonised information,
  - Enabling to visualise the complete history of a substance across processes
    - Stop Excel tables to 'fill the gaps' of information that was not recorded in the first place
- Give good data to ECHA Committees and MSCAs
  - To inform opinions
  - To make the ground for such opinions fully traceable
- **Integrate data collected in the first era with data of the second era into a "substance universe"**

Mapping The Universe

Mapping the EU chemical universe is understood as a categorisation of registered substances into “bins”



1. High priority for further work by authorities

Substances with identified concern and regulatory work has already been initiated or further work can be initiated based on currently available data

2. Substances of unknown priority

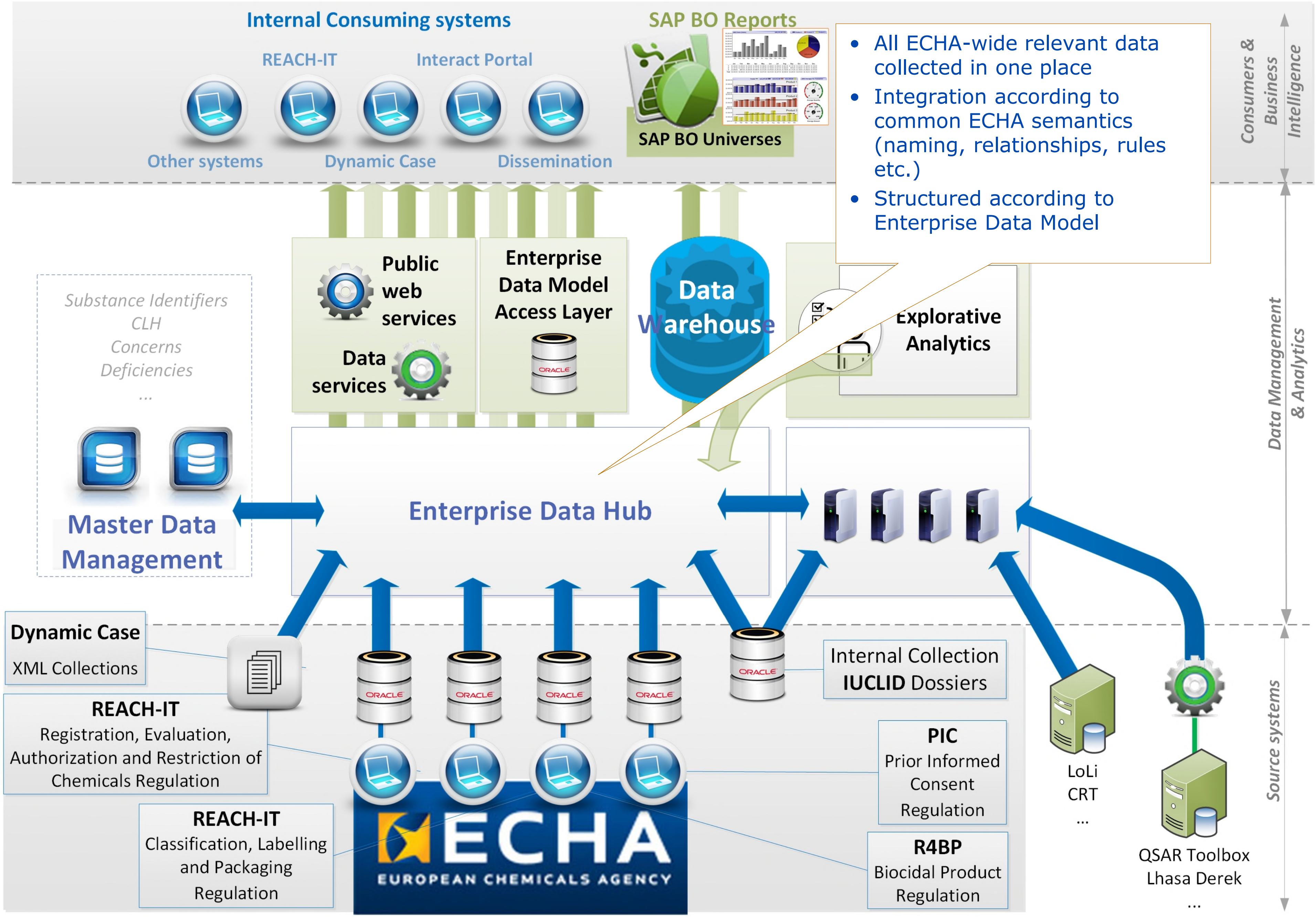
Substances for which there is at present uncertainty regarding the hazardous properties and/or the potential exposure; risk cannot be excluded although it cannot be established based on currently available data.

3. Low priority for further work by authorities

Substances for which available data suggest that no further regulatory action is needed at present

Let Me Be An IT Manager for one slide

ECHA IT Enterprise Architecture
for data managements
and scientific data analytics



- All ECHA-wide relevant data collected in one place
- Integration according to common ECHA semantics (naming, relationships, rules etc.)
- Structured according to Enterprise Data Model

Consumers & Business Intelligence

Internal Consuming systems

REACH-IT

Interact Portal

Other systems

Dynamic Case

Dissemination

SAP BO Reports

SAP BO Universes

Data Management & Analytics

Public web services

Data services

Enterprise Data Model Access Layer

Data Warehouse

Explorative Analytics

Substance Identifiers

CLH

Concerns

Deficiencies

...

Master Data Management

Enterprise Data Hub

Source systems

Dynamic Case

XML Collections

REACH-IT

Registration, Evaluation, Authorization and Restriction of Chemicals Regulation

REACH-IT

Classification, Labelling and Packaging Regulation

Internal Collection

IUCLID Dossiers

PIC

Prior Informed Consent Regulation

R4BP

Biocidal Product Regulation

LoLi

CRT

...

QSAR Toolbox

Lhasa Derek

...

ECHA

EUROPEAN CHEMICALS AGENCY

Benefits Of Structuring Information

- Easier to **identify**, from a set of defined fields, what **key information** is expected to be submitted within a specific regulatory context
- Possibility to **format the data** automatically (e.g. assessment reports)
- **Search** possibilities are increased allowing data mining and prioritisation
- Existing data stored in a Harmonised Template can be processed in order to **prepare data submissions** to answer different regulatory requirements
- **Exchange** of data between different IT tools is facilitated
- **Validation** before submission is facilitated
- Regulatory work, dissemination is much more **efficient**
- **Consistency, single semantic, traceability, interoperability**
- **Analytics**
- **… Managing knowledge**

Benefit Of The Doubt

Is from *unstructured* to *structured* always good?

It depends!

For example

Weak signals of an effect, scattered over several studies can defeat the point

Other means, such as text analytics, could be better suited…

If one can appreciate the difference in what she gets out of it

EFSA: What are we doing in the IUCLID pilot

- Learning from biocides – creation of submission types for active substances, micro-organisms and products
- Analysing business rules from REACH – to automate technical dossier checks
- Developing reports with IUCLID data – supporting risk assessment

EFSA: What do we need to think about

- Capturing and integrating data created during the risk assessment process
    - Critical appraisal tools
    - Validated endpoints
    - Outputs from tools, models and calculators e.g. BMD, PRIMO, OPEX

Thank you!

luisa.consolini@echa.europa.eu