Training Seminar on Data and Analysis in Environmental Audit

Tirana, 17.10.2017

Please discuss within your group, what kind of problems / difficulties have you faced:

- with data gathering and data availability;

- with data quality and preparing data for analysis;

- with merging different data(sets), using data from different sources

- with data analysis

What have been your greatest successes with data collection and data analysis in auditing process

Environmental Data: Resources and Options for Supreme Audit Institutions

INTOSAI WGA 2013

Environmental Data: Resources and Options for Supreme Audit Institutions (INTOSAI WGA 2013)

The research paper discusses the main ways that SAIs use environmental data:

- some **key sources of environmental data** that are available to audit institutions at the global, regional, and other levels,

- as well as **key considerations when using such data**; and

- **a variety of tools and methods** that audit institutions can use **when high-quality environmental data are lacking.**

Definition of Environmental Data

Environmental data as **systematically collected qualitative or quantitative information** about different components of the environment or human activities and sectors that affect the environment.

Environmental data may undergo several steps of processing, depending on the information needs of the user.

Insufficient data, lack of high-quality data

When high-quality environmental data are not available, options still exist:

- use related data to estimate unavailable data,

- comparing several data sources for estimations,

- develop their own data, or

- asking voluntary input to the audits.

- SAI may also decide that the absence of high-quality data will be a central audit message.

Collecting own data

SAIs can **collect their own data** to meet the audit's needs through questionnaires, surveys, or physical observations from site visits or photos.

There is need to consider

- how cost-effective the work will be,

- what impact the data will have on the audit work and

- whether the auditing team/SAI has the capacity or expertise to develop the data (incl. quality issues).

- whether they can develop data covering all the relevant areas,

- whether they would need to use a sample that would be statistically representative, or

- whether their data and possibly conclusions, would be more limited.

Data quality assessment

SAIs should consider assessing the **sufficiency (quantity) and appropriateness (quality) of data for use in an audit.**

SAIs can undertake a variety of activities to conduct a data assessment, such as:

- reviewing existing information about the data, such as external data assessments, reports, or studies;

- conducting interview with officials responsible for managing the data to determine such things as the quality of the underlying data (metadata), and how and when officials entered the data into databases or records management systems;

- performing tests on the data to identify the extent of inconsistencies, completeness, or errors;

- tracing data in computer systems to and from source documents; and

- reviewing selected internal quality controls for data or records management systems.

Data quality issues

- **Relevance** reflects the degree to **which the information meets the needs or requirements of clients, users, stakeholders or the audience**.

- **Accuracy and reliability.** Accuracy reflects the degree to which **the information correctly describes the phenomenon** it was designed to measure. Reliability concerns whether the data consistently over time measure the reality that they are designed to represent.

- **Timeliness and punctuality.** Timeliness refers to **how fast – after the reference date or the end of the reference period – the data are released or made available**. Punctuality refers to **whether data are delivered on the dates promised, advertised, or announced.**

Data quality issues

- **Accessibility and clarity.** The **data and metadata can be found or obtained without difficulty, are presented clearly, are available and accessible to all users** on an impartial and equal basis.

- **Coherence and comparability.** The data are **consistent internally and over time and are produced using common standards with respect to scope, definitions, classifications and units.** Users should to be able to combine and make joint use of related data from different databases

AUDITING PROCESS

Auditing process and need for data

Monitoring	Planning	Conducting audit	Reporting
Monitoring, Looking for the audit topic, Importance of the topic for auditing	Selecting topics and auditing questions; Deciding research design; Selecting data and data analysis methods; Selecting samples or case studies	Data collection and analysis, Answering the research questions; Are all aspects covered and all findings are supported by facts?	Providing context for audit findings and recommenda-tions

AUDIT / RESEARCH QUESTIONS

A "What, Why, and How" framework

What?	Why?
What do I want to know more about / understand better? What are my key research questions?	Why will this be of enough interest to others? Can the research be justified as a 'contribution to knowledge'?
How – conceptually?	How – practically?
What concepts (models, theories) can I use to answer my audit/ research questions? How can these be brought together into a basic conceptual framework to guide my investigation?	What investigative styles and techniques shall I use to apply my conceptual framework (both to gather material and analyse it)? How shall I gain and maintain access to information sources?

Audit / research questions

The essential **criteria for evaluating research questions include their soundness and clarity**, but also **whether they can be answered in the framework of given and limited resources** (time, money, etc.).

Clear ideas about the nature of the research questions that are pursued are also necessary **for checking the appropriateness of methodological decisions** in the following respects:

- Which methods are necessary to answer the questions?

- Is it possible to study the research question with the chosen methods at all?

Research Questions

- guide your decisions about the kind of research design to employ,
- guide your decisions about what data to collect and from whom,
- guide your analysis of your data,
- guide your writing up of your data,
- stop you from going off in unnecessary directions and tangents.

CONSTRAINTS ON RESEARCH QUESTIONS

- We cannot answer all the research questions that occur to us.
- We therefore have to select from the possible research questions that we arrive at.
- We should be guided by the principle that the research questions we choose should be related to one another.

Preparing for your research

- **Do not begin your data collection until you have identified your research questions reasonably clearly.**

- **Develop your data collection instruments with these research questions at the forefront of your thinking**
  - If you do not do this, there is the risk that your results will not allow you to illuminate the research questions

- **If at all possible, conduct a pilot study to determine how well your research instruments work**

It is important, that the researcher develops a clear idea of his/her research question, but **remains open to new and perhaps surprising results.**

DATA AND DATA COLLECTION

Data – different possibilities for classification

- National, regional, global, spatial data (WP 2013)
- Qualitative and quantitative data
- Verbal, visual, textual data
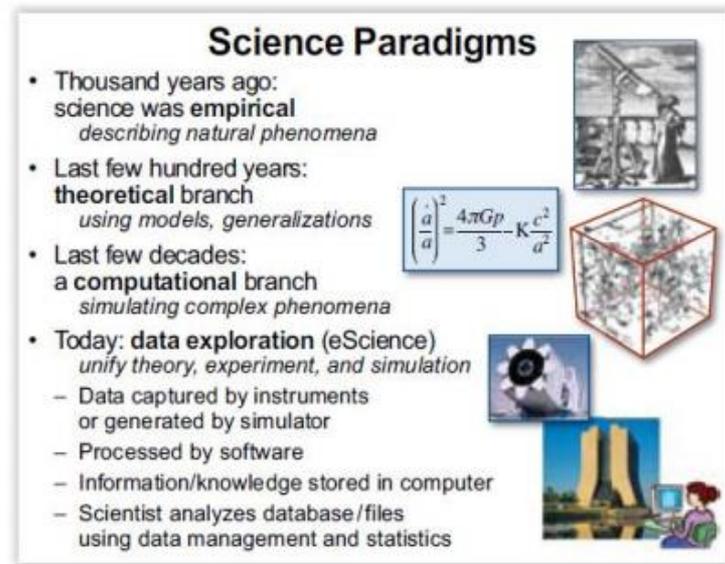- Primary and secondary data

- Big data
- Open data

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$

- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

Figure 4. Science Paradigms from Hey et al. (2009).

Complementary information and tools related to data sources

It may be helpful to **consult the complementary information** – including data documentation and other tools (e.g. direct support from tutorials, and other guidance):

- Data documentation can tell **what the data represents – what the data measure, what measurement units are used, and how the units and key categories are defined.**

- Documentation of methods can help **to identify assumptions, quality controls and standards, and other aspects of the data set that may be critical for setting limits on the scope and conclusions of the audits.** Thorough data documentation will also include a discussion of key uncertainties and their origins.

- **The techniques used to collect data may also affect how the data can be used** (voluntary reporting and problems with underestimation; automatic measurements and need that measurement instruments are calibrated ad standardised).

Data – different possibilities for classification

- National, regional, global, spatial data (WP 2013)
- Qualitative and quantitative data
- Verbal, visual, textual data
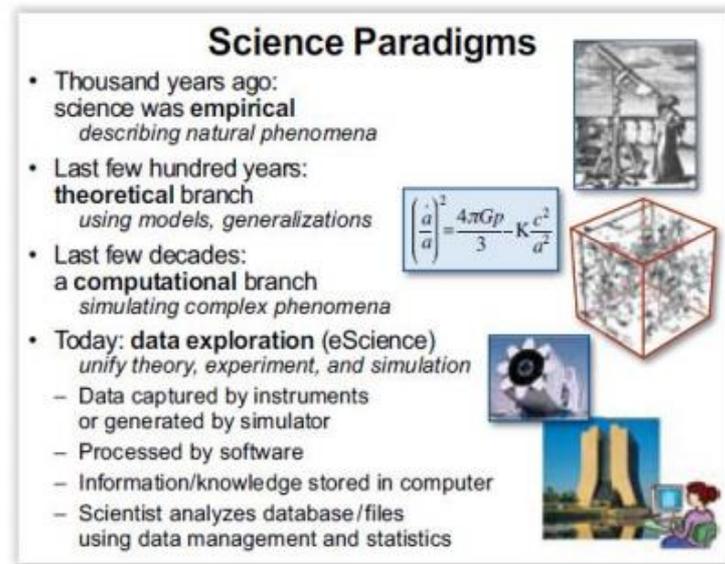- Primary and secondary data

- Big data
- Open data

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$

- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

Figure 4. Science Paradigms from Hey et al. (2009).

Big Data

Big data can be described by the following characteristics:

- **Volume** - the quantity of generated and stored data
- **Variety** - the type and nature of the data.
- **Velocity** - the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
- **Variability** - inconsistency of the data set can hamper processes to handle and manage it.
- **Veracity** - the data quality of captured data can vary greatly, affecting the accurate analysis.

Open Data

Open Data Portal of Estonia

The Open Data Portal provides a single point of access for general public to unrestricted public sector data with the permission to re-use and redistribute such data for both commercial and non-commercial purposes.

Search for data...   **SEARCH**

Goals of opening up data

- ◆ to stimulate economy
- ◆ to increase transparency
- ◆ to provide access to information
- ◆ to decrease the number of information requests and, consequently, also reduce the workload of public sector workers
- ◆ to facilitate creation and management of open services for private and community sectors
- ◆ to encourage migration to future technologies such as Linked Data, Big Data and Internet of Things

Research question, data and data analysis

Research questions:

- **Descriptive** – usually we want to describe something (e.g. status of something or some kind of process using words/text or having numerical data).

- **Causal** (research questions dealing with causality). There is always question, how well we can analyse the causality and whether there exists data, which allow us to interpret the results in the perspective of causality.

- **Predictions** (e.g. prediction of energy use, oil price, etc.).

AUDIT QUESTIONS, DATA AND DATA ANALYSIS

THANK YOU!

Kaia Philips
kaia.Philips@riigikontroll.ee