# Impact assessment for creating more good: Data-centric led approach

Christopher Ibeh
GeoDigital Sustainability Group

## Abstract

Beyond the 'do no harm' principle, Impact Assessment (IA) can be a tool that promotes transformational change towards environmental sustainability as well as low carbon and climate resilient development. There is the need for approaches to make IA fit for purpose, taking IA from the do less harm to the creating more good principle. On the other hand, in a bid to stimulate economic growth and speed up decision making, approaches such as the simplification or streamlining of IA has been promoted in many countries. The streamlining of IA may result in sacrificing its essence and principles. This paper focus on data-centric approach for biocentric IA outcome that is fit for purpose. The beauty of digitising is not only in creating more data and generating more value from data, but also ensuring the quality of data. Creating high quality data is a pre-requisite for the compounding effect of moving beyond just risk-based IA to creating more good.

Key words: Data Centric , Nature positive, Do more good, Environmental impac6 assessment, Artificial Intelligence, Exponential Technologies, Data quality.

## Introduction

IAIA has existed since the 1980's and over the years we have improved on tailored methodologies for impact assessment, such as Cumulative impact assessment, Social impact assessment, healthy impact assessment, ecological impact assessment among others. However one area where we have not consciously pushed boundaries is in data centric Impact assessment for nature positive solution. There is the need to move from not just data driven IA to data centric Impact assessment while also maintaining data driven approaches to impact assessment.

This study aims to demonstrate the need for a data centric approach to impact assessment (IA) for nature positive solution. Specifically, this study had a three-fold objective: 1) to show the importance of data centric IA for nature positive solution, ii) to provide a data centricity framework, and iii) use the data centricity framework to evaluate a biodiversity data aggregator and provide workable solution where necessary. The discussions and evidence presented here can underpin or inspire further studies and policies in a variety of contexts.

The paper is divided into six sections, besides this introduction. The next section provides a background on the existence of data quality spectrum and the importance of data quality. Section three explains the relationship between exponential technologies such as artificial intelligence (AI) and impact assessment (IA). Section

four explains the link between data centricity and nature positive IA. Section five presents and discusses a data centricity framework developed as part of this study and section six present a case study applying the data centricity framework. Finally, section seven presents concluding remarks.

## Data quality spectrum

Considering that EIA is about the future and as such is based on imperfect knowledge, it may not be possible to have a perfect understanding of the environment. The use of data of good quality is therefore critical to the usefulness of impact assessment predictions. This has become much more important with the rise of exponential technologies like AI which has the potential to generate data (i.e., generative AI) which could be erroneous if the model was trained with poor quality data.

Figure 1a shows the proportion of occurrence of the word "data quality" in EIA articles on google scholar for the past 3 decades. The plot shows that “data quality” mention proportion significantly increased since around the year 2019. This imply that data quality has recently become more topical in EIA papers. Figure 1b shows the percentage change in the measured proportion of mention of the word “data quality” in EIA articles on Google scholar, and this buttresses the point of an increasing trend. Similarly, Figure 2a shows the proportion of occurrence of “nature positive” in environmental impact assessment articles. The plot shows a stagnating trend up until about 2020 where it started to increase significantly. This also shows that “nature positive” concept have become an increasingly important concept in environmental impact assessment. This has become necessary due to the need for impact assessment for more good over the do less harm ideology.

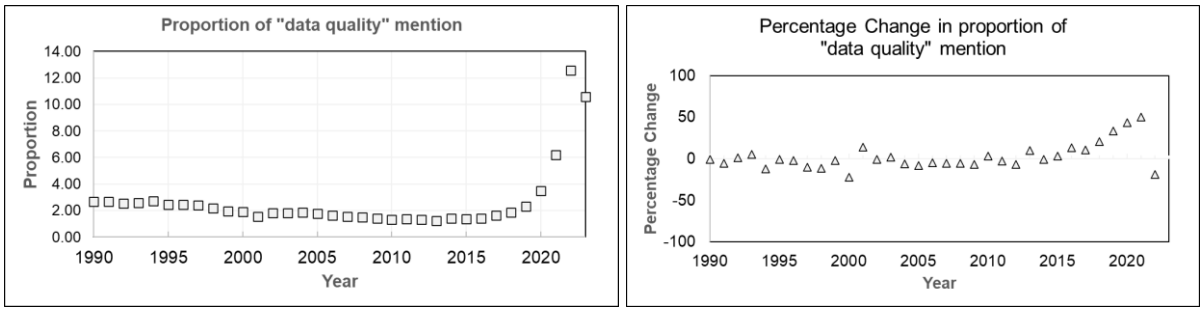


Figure 1: a) plot of proportion of “data quality” mention in environmental impact assessment articles against year form 1990 to 2023 b) plot of percentage change in proportion of “data quality” mention in environmental impact assessment articles against year form 1990 to 2023 (data retrieved from Google Scholar).

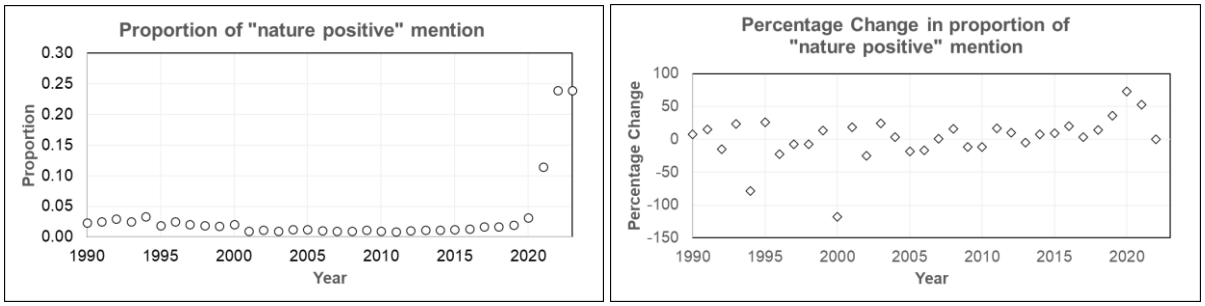


Figure 2: a) plot of proportion of "nature positive" mention in environmental impact assessment articles against year form 1990 to 2023 b) plot of percentage change in proportion of "nature positive" mention in environmental impact assessment articles against year form 1990 to 2023 (data retrieved from Google Scholar).

Data exist on a broad spectrum. On one end is perfect quality data/certitude for impact evaluation, and on the other end is non existence of data or totally erroneous data, and yet in between these two ends are some degree of poor/partial/limited quality data where a total reliance on risk assessment/uncertainty analysis is required. The goal is to achieve a perfect state where we have sufficient knowledge of environmental systems and as such be able to state with certainty the environmental impact of projects. Although this may not be the case at this time, however, AI and the drive towards singularity creates an opportunity for us to move closer towards that perfect information stage. This is also linked to nature positive concept in that we need quality data to determine with certainty that nature is enhanced from the current state. This is the basis of impact assessment for more good rather than less harm because less harm is majorly due to the existence of uncertainty and hence the need to reduce possible harm.

## Exponential technologies and impact assessment

Digital impact assessment especially the introduction of Artificial intelligence (AI) into environmental impact assessment is inevitable because AI is an exponential technology which are in their nature are deceptive and destructive (see figure 3) just like the Agricultural revolution, the industrial revolution and the internet/information revolution. This has become more important as we are entering an age of the maturation of multiple exponential technologies (Artificial Intelligence, Synthetic Biology, Nano Technology, Quantum Computing, Blockchain Technology), see figure 4. AI will affect all industries and sectors of the economy and environmental Impact assessment will not be an exception.

To ensure that IA is fully integrated and to unlock the potentials of AI for Impact assessment we need to ensure the availability of quality data as AI result is quality data dependent. The advent of generative AI has made it more important than ever to ensure a data centric Impact Assessment. This is important in that generative AI creates new data by learning from existing data. Learning from poor quality data means generating poor quality data.

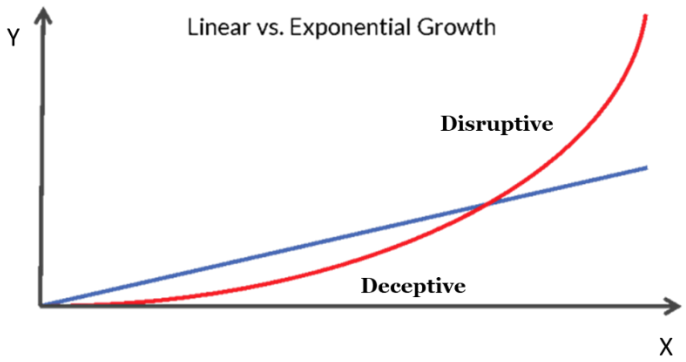


Figure 3 : Exponential technologies such as Artificial intelligence are deceptive and destructive. They will affect about all industries hence we need to pay attention to them

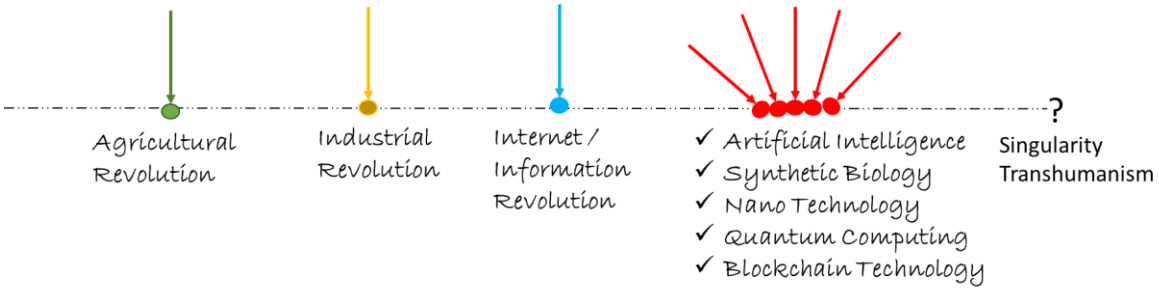


Figure 4 : The maturation of multiple exponential technologies at the same time require that data centricity is the foundation of impact assessment.

## Data centricity and Nature positive impact assessment

Focussing on data centric and nature positive impact assessment will unlock impact assessment for more good. Data drivenness on one hand means acting on data while data centricity means a culture that treats data as an asset and hence ensure it is of good quality, ready for use. Nature positive is a world where nature – species and ecosystems - is being restored and is regenerating rather than declining (WEF, 2021). Data centricity becomes critical towards achieving nature positivity because if we don't have quality data on the baseline of species and ecosystems then we cannot adequately measure progress or the reverse. There is the need to move from not just data driven IA to data centric Impact assessment while also maintaining data driven approaches to impact assessment.

## Data centricity framework

There are 4 categories of strategies for achieving data centric environmental impact assessment. Data cleaning, data management, data validation and data life cycle management. This can include creating a centralized database or repository for storing data, as well as developing protocols for how data should be collected and stored. Data cleaning encompasses the process of identifying and correcting errors or inconsistencies in the data, and standardizing the data for consistent across different

sources. Data management involves organizing and storing data in a way that makes it easy to access, retrieve, and analyse. Data validation involves checking accuracy and consistency of the data and Data life cycle management involves the documentation of data provenance as it moves from phase to phase. Bases on these strategies, a framework is hereby presented for evaluating data centricity. The framework is shown in figure 5.

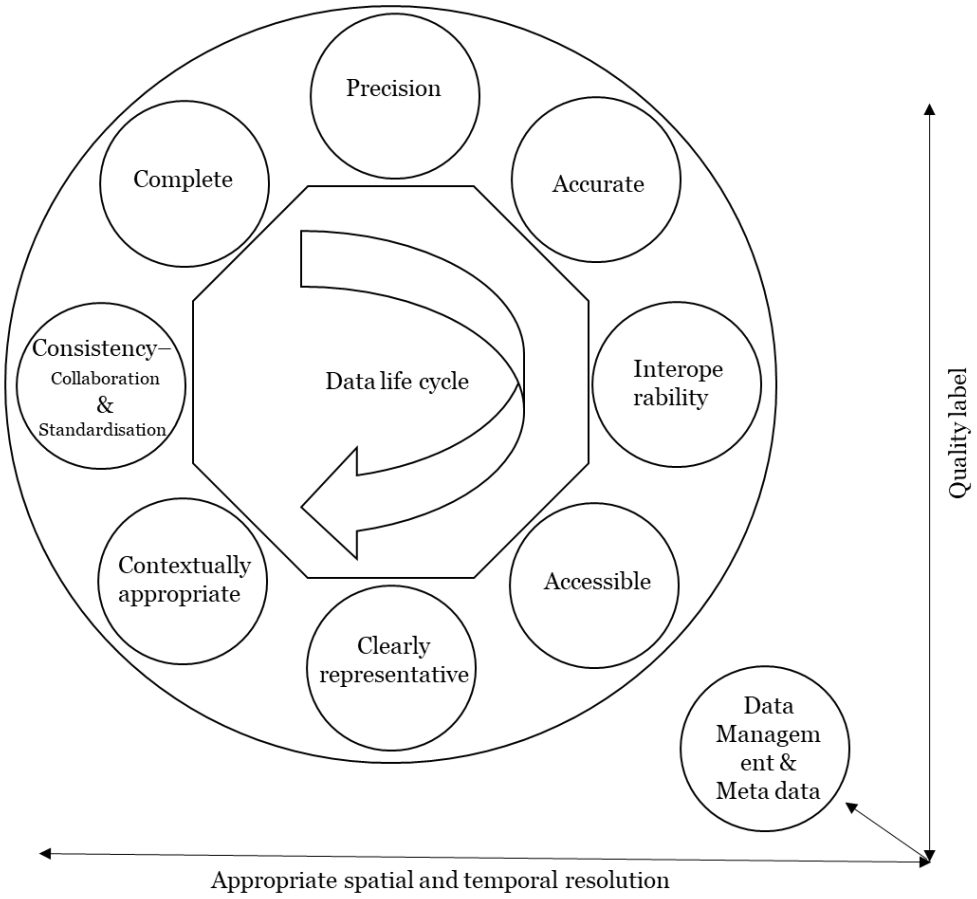


Figure 5: Data centricity framework

These include :

i. Accurate – this is the state of being correct. It answers the question, is the data representative of the real world.
ii. Interoperability- a characteristic of a data being able to work with other data.
iii. Consistency and reliability through collaboration and standardisation - all records of data should be the same, regardless of its sources or movements across systems. All forms of data transformation the data has undergone should also be indicated.
iv. Contextually appropriate and relevant - relating to the circumstances that form the setting for the data.

v. Clearly representative – data should portray similar characteristics as the population
vi. Accessible - the ability of people to be able to both see and use the data.
vii. Appropriate spatial and temporal resolution (and Timeliness)- Data is usually represented in space and time, the data scale has to be the required scale.
viii. Completeness- the comprehensiveness or wholeness of the data. There should be no gaps or missing information for data to be truly complete
ix. Quality label- adding quality tags or labels to raw data. This ensures better quality assurance
x. Data life cycle -this reflects data provenance and Reflect post Phase changes
xi. Precision—Does the data fall within the range of acceptable values
xii. Data Management and Meta data. There should be a management structure in place for the data. Meta Data is the data about the data. Every data should have a meta data containing full description of the data.

Ensuring good data quality requires a combination of careful data collection and management practices, as well as robust quality control procedures.

Overall, Impact assessment has done well on data accuracy, and representativeness dimensions for example . However Impact assessment has not done well on data accessibility, interoperability and quality label dimensions- For example at province level IA is conducted in a different way with different thresholds set by politically agreed thresholds and as such it is challenging the harmonisation of information (interoperability). There is also the issue of data hoarding, countries and companies and project owners hold on to their data,. As such the community cannot necessarily pull data resources together easily.

**CASES STUDY:**

**Global Biodiversity Information Facility (GBIF) Data**

GBIF is an international network and data infrastructure funded by the world's governments and is aimed at providing anyone, anywhere, open access to data about all types of life on Earth. Its vision is to create a world in which the best possible biodiversity data underpins research, policy and decisions. Moudrý and Devillers (2020) in their paper title “Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data” evaluated the potential information gaps in marine mammal distributions using data from data aggregators such as GBIF and Ocean Biodiversity Information System (OBIS). They did so by overlaying International Union for Conservation of Nature (IUCN) range maps and species occurrences from global databases and found that areas previously identified as hotspots for marine mammals' diversity show some of the highest rates of potential false positives (i.e. species are thought to occur there based on their range map, but no species record exist in either GBIF or OBIS). Their study pointed to data quality challenges that can limit data usability in biodiversity research.

Evaluating GBIF data with the data centricity framework shows that although it does well in many areas it is however has not done well in terms of consistency which can be achieved through collaboration possibly with other collaborators and the provision of data quality label. GBIF can make their data more usable for nature positive endeavour by adding quality tags to data.

We can learn from medical science where data may be the deciding factor between life and death and data is restricted by privacy. GBIF can apply an Automated detection of poor-quality data through the application of data quality labelling techniques as suggested by Dakka et al (2021). This will enable GBIF to better achieve their vision of making available best possible biodiversity data.

For environmental impact assessment, it is important to have a centralised collection of environmental impact assessment statements and data. The data with which each outcome were reached all harmonised with the data centricity framework. Although there are limitations of privacy, intellectual property right etc, the work at H2S new high-speed railway project in the United kingdom has shown that it is possible for companies to collaborate and share data in this regard. With the enrichment of such data, it can provide an avenue for impact assessment professionals to validate the data, it will provide transparency and an opportunity to see the bigger picture when data puzzles are put together. Such centralised system will also provide an avenue for the verification of information especially as we enter the age of AI.

## Conclusion

Data centric Impact Assessment is required to move IA from the idea of do less harm to the idea of do more good. It is essentially the foundation of nature positive solutions. This has become much more important as we prepare impact assessment for the age of Artificial intelligence and the move towards singularity.

The data quality framework developed in this study can act as a data quality evaluation tool. I am currently working on developing a tool using transfer learning to help label data quality first for biodiversity data and subsequently other data types.

Reference

Dakka, M.A., Nguyen, T.V., Hall, J.M.M., Diakiw, S.M., VerMilyea, M., Linke, R., Perugini, M. and Perugini, D., 2021. Automated detection of poor-quality data: case studies in healthcare. Scientific Reports, 11(1), p.18005.

Moudrý, V. and Devillers, R., 2020. Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. Ecological Informatics, 56, p.101051.

https://www.weforum.org/agenda/2021/06/what-is-nature-positive-and-why-is-it-the-key-to-our-future/(Accessed on 20/02/2023)