Archival Odyssey: Part 2

Dr Anthea Seles, Secretary General
8 April 2019
Norwegian Triennial Archival Conference
Archival Considerations

Impact of artificial intelligence, machine learning and data mining in government

How we use artificial intelligence in archival processes

Making records accessible and readable for research
Government Use of Artificial Intelligence and Machine Learning

• Decisions are being made now using machine-learning and artificial intelligence

• For example, data science or statistical analysis units in government departments and private corporations
  • Data science and the ability to mine data is seen as a competitive advantage.
  • Platforms that is common usage use these techniques: Netflix, Google, Facebook etc
  • For government it is seen as a way to parse through large volumes of data (structured and unstructured) to make a decision
  • Visualisations for policy decisions
Government Use of Artificial Intelligence and Machine Learning

• There are challenges with the data science approach and the use of machine-learning and AI algorithms in government decision-making:
  • Is the data we are combining meant to be combined? Are we simply comparing apples and oranges?
  • Is the data biased and how does that affect the output of the algorithm? How does that affect what we see and how we interpret it?

• Archivists have often played a role in advising organisations on the creation and preservation of records and data to ensure their evidentiary value:
  • What advice would we give in the creation and preservation of ‘algorithmic records’?
  • Does the archivist have a role to play in advising how algorithms and code are created for decisions-making? How do we know what to preserve and how?
Government Use of Artificial Intelligence and Machine Learning
Considerations:

• If this becomes standard practice in government and passes into policy how do we begin to advise on what documentation needs to exist to document the training data and subsequent information that is input or not into the system? What does integrity and accountability look like in this context? By extension, what do we preserve?

• Does the archivist have a role as an ethical advisor in this context?

• To read the article: https://news.sky.com/story/handwriting-to-help-govt-catch-gangs-behind-mass-scale-benefit-fraud-11190448
Example:

- Cathy O’Neil *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*
- In some US states they use algorithms to help determine recidivism rates (COMPAS - Correctional Offender Management Profiling for Alternative Sanctions)
- Some context of the data that was used to train COMPAS the algorithm created by Northpoint
  - Sentences given to African-American prisoners in the federal system is 20% longer than those given to white convicts for similar crimes
  - African-American represent 13% of the population of the United States, but account for 40% of the prison population
- Base training data set is biased and then the algorithm is created by a private company, which makes it a black box
Artificial Intelligence and Machine Learning in Archival Processes


• The focus of the discussion will be on: appraisal and selection, and identification of sensitive information in record sets

• There are large volumes of unstructured records sitting in file shares or defunct and legacy systems all over the world

• The ability of archivists to have a definable and useful entity such as a folder, to appraise and select is not always a certainty
Artificial Intelligence and Machine Learning in Archival Processes

• Volume also complicates the ability of archivists to be able to assess at scale unstructured record sets.
  • For every 1TB in information management systems: ~25TB in file shares or other unstructured records environment. This does not account for any datasets or contents of email servers.
  • Once datasets and email servers are accounted for you can be looking at upwards of 1.5PB of records that you will need appraise and select. 1.5PB = approx. 1.5 billion word documents.
  • This information can also contain varying levels of context and limited metadata. The metadata can also be inaccurate because of previous records migrations.

• Automation is no longer a choice, it is a necessity but that does NOT mean the archivist (the human) is irrelevant in the process.
Artificial Intelligence and Machine Learning in Archival Processes

• The challenge with automating appraisal and selection, along with the sensitivity review process:
  • How do you measure accuracy? What does ‘good enough’ look like? What are the risks? What is acceptable risk appetite?
  • How can we determine what might be missing?
  • How can be accountable for the decisions we make based on machine outputs? How do we equally hold the machines to account?
  • How do we compensate for the change in the digital record over time? Retune the algorithm?

• We are dealing with ‘Black Boxes’

• RISK: Biasing the historical record and by proxy history and our collective memory
Artificial Intelligence and Machine Learning in Research

- Two issues for the archival community to consider:
  - Impact of researchers trying to mine archival data
  - Digitisation of historical data and information

- Researchers are starting to use data mining techniques to parse through large volumes of digital data.
  - Ex: Researchers are using tools like Google NGRAM to mine literature to trace things like stereotypes in literature

- There are also many other tools, sometimes bespoke, that researchers are or will begin using.
Artificial Intelligence and Machine Learning in Research

• There is a question for archivists about how much access we may wish to allow researchers access to public records and data
  • Data mining and machine learning tools breakdown silos created by archival description (i.e. fonds, series, files)
  • Can reveal unknown connection that become sensitive or problematic by virtue of making that connection
  • Can surface sensitive information that was missed during sensitivity review
  • Also once the data is mined and put into a system outside the archives, what else can it can be combined to?
We also need to consider the impact of future digitisation.

The re-purposing and re-use of archival records and data has enormous value and I think we sacrificed much of digitisation and allowing companies to digitize archival records and data, in order that we can get a ‘free’ copy. We must be more savvy.

Companies are beginning to realise the value of data held in historical records. Digitising them and applying OCR is a method for gaining access to large volumes of data to train algorithms.

We need to start asking ourselves:

- Why is the digitisation free?
- Will this data be used to train an algorithm?
- What is the company’s ethical stance?
- What happens to the data once the digitisation is done?
- Will there be an impact on people’s lives?

Scenario: Paper death registrations
Conclusion

• Government Use of Artificial Intelligence:
  • What role does the archives and information communities have to play in this space? Do we have a role?
  • What skills do we have or do we need if we have a role to play?
  • What is the ‘record’? How do we capture and preserve that record?
  • Who are our partners? How do we begin to work with them?

• Machine Learning and Artificial Intelligence in Archival Processes
  • What is accuracy? What risks are we willing to accept?
  • How can we ensure the accountability of the decision we make based on machine-learning and AI processes?

• Artificial Intelligence and Machine Learning in Research
  • How much access is too much when machines are involved?
  • What are the right questions to ask when private companies offer us free digitisation?
  • How do researchers want to use our records to carry out digital research?
Whether you are using an algorithm, artificial intelligence, or machine learning, one thing is certain: If the data being used is flawed, then the insights and information will be flawed.

-Venkatesan M Artificial Intelligence vs Machine Learning vs Deep Learning
References and Further Reading


References and Further Reading

Thank you.

Dr Anthea Seles
Secretary General
International Council on Archives
seles@ica.org