International Council on Archives
Conseil International des Archives

Archival Odyssey: Part 2

Dr Anthea Seles, Secretary General

8 April 2019

Norwegian Triennial Archival Conference

International Council on Archives
Conseil International des Archives

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

International Council on Archives
Conseil International des Archives

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

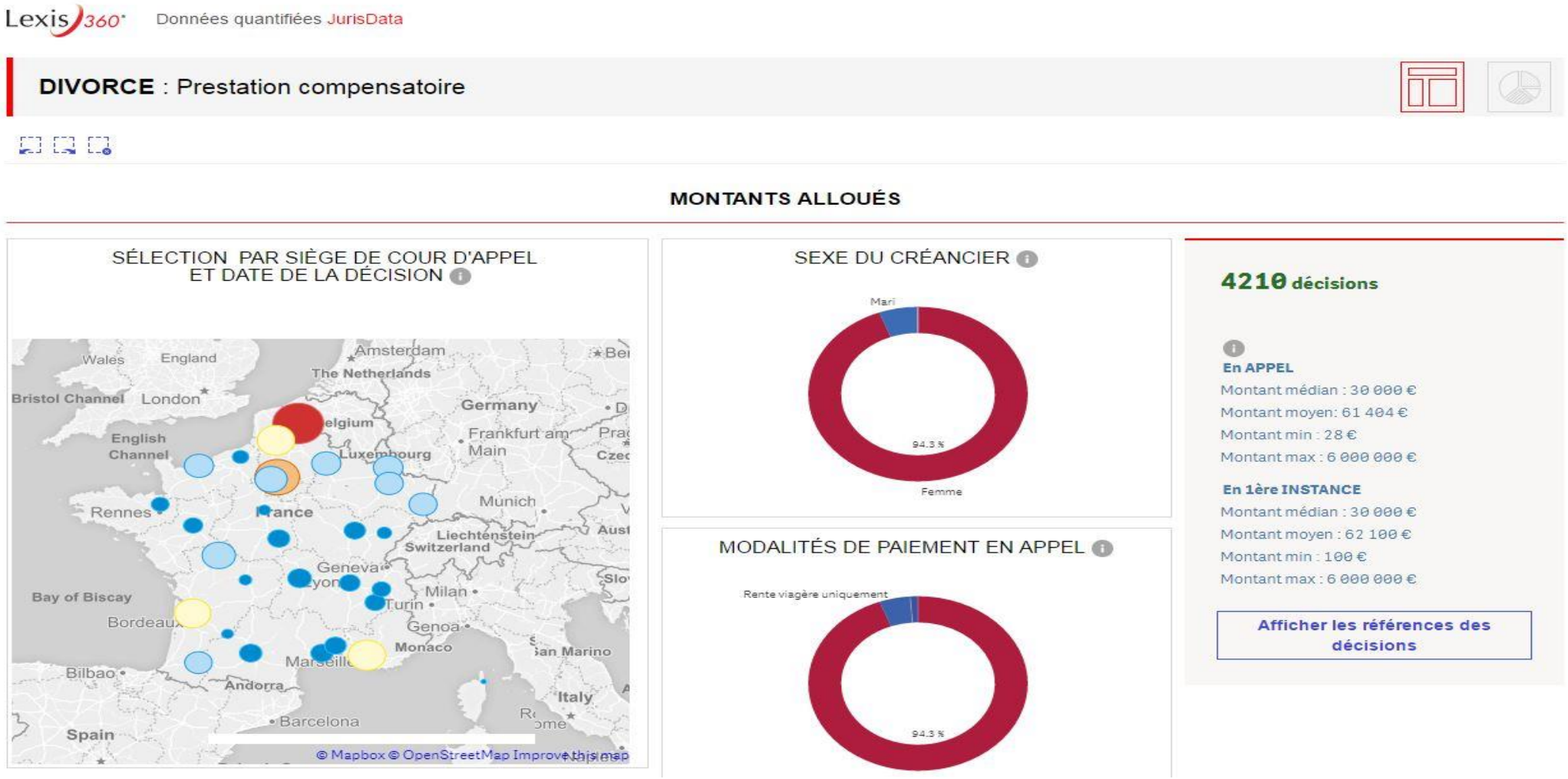
International Council on Archives
Conseil International des Archives

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?

International Council on Archives
Conseil International des Archives

Government Use of Artifical Intelligence and Machine Learning



International Council on Archives
Conseil International des Archives

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

sky news

Search

Watch Live

Home UK World Politics US Ocean Rescue Science & Tech Business Ents & Arts Offbeat Weather

Handwriting to help Govt catch gangs behind mass-scale benefit fraud

Artificial intelligence is going to be used to clamp down on cheats claiming bogus benefit payments worth millions of pounds.

19:31, UK,
Sunday 31 December 2017

A record £1.1bn in overpaid benefits was recovered from fraudsters last year

International Council on Archives
Conseil International des Archives

Government Use of Artificial Intelligence and Machine Learning

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

International Council on Archives
Conseil International des Archives

Artificial Intelligence and Machine Learning in Archival Processes

- *The Application of Technology Assisted Review to Born-Digital Records Transfers, Inquiries and Beyond*. (2016) London: The National Archives UK http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf
- 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

International Council on Archives
Conseil International des Archives

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 contains 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

International Council on Archives
Conseil International des Archives

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? Re-tune the algorithm?
- We are dealing with ‘Black Boxes’
- RISK: Biasing the historical record and by proxy history and our collective memory

International Council on Archives
Conseil International des Archives

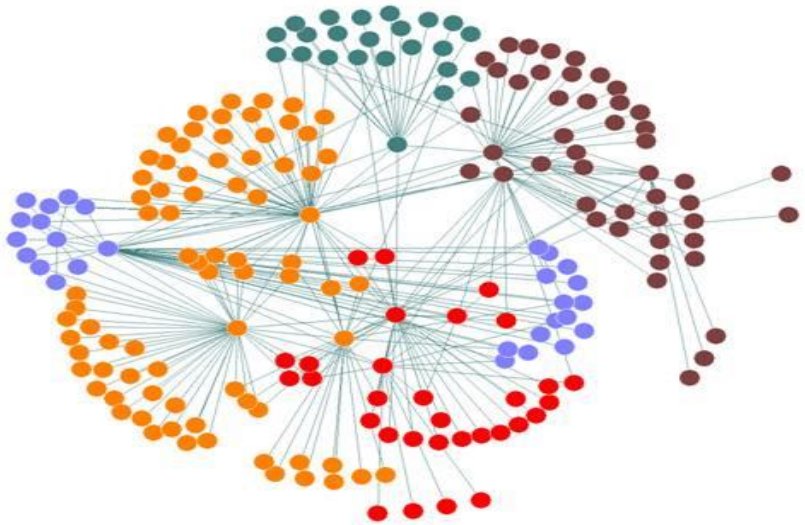
Artificial Intelligence and Machine Learning in Research

- 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
    - Susan Mason. 'Analysing Stereotypes Across Time Using Google Ngram Viewer' *SAGE Research Methods Cases Part 2* (2018) doi:10.4135/9781526436245
- There are also many other tools, sometimes bespoke, that researchers are or will begin using.

International Council on Archives
Conseil International des Archives

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 siloes 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?



International Council on Archives
Conseil International des Archives

Artificial Intelligence and Machine Learning in Research

- 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

International Council on Archives
Conseil International des Archives

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?

International Council on Archives
Conseil International des Archives

A parting tought…

Whether you are using an algorithm, artificial intelligence, or machine learning, one thing is certain: If the data being used if flawed, then the insights and information will be flawed.

-Venkatesan M *Artificial Intelligence vs Machine Learning vs Deep Learning*

References and Further Reading

- *The Application of Technology Assisted Review to Born-Digital Records Transfers, Inquiries and Beyond*. (2016) London:The National Archives UK http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf
- Bhaskar, Michael. *Curation: The Power of Selection in a World of Excess.* (2017) London:Piatkus
- Caplan,Robyn, Joan Donovan, Lauren Hanson and Jeanna Matthews. 'Algorithmic Accountability: A Primer' *Data and Society* (2018) https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf
- Chumtong, Jason and David Kaldewey. 'Beyond the Google NGRAM Viewer: Bibliographic Databases and Journal Archives As Tools for Quantitative Analysis of Scientific and Meta-Scientific Concepts. *FIW Working Paper No 8* (2017) https://www.fiw.uni-bonn.de/publikationen/FIWWorkingPaper/fiw-working-paper-no.-8
- Delort, Pierre. *Le Big Data* (2015) Paris : PUF
- Domingos, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (2015) New York:Basic Books
- Engin, Zeynep and Philip Treleaven. 'Algorithmic Government: Automating Public Services and Supporting Civl Sevants in using Data Science Technologies' *The British Computer Society* (August 2018)https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxy082/5070384
- Ertzscheid, Oliver. *L'appétit des géants: pouvoir des algorithmes, ambitions des plateformes* (2017) Paris : C&F
- Information Privacy Commissioner. *Big Data, Artificial Intelligence, Machine Learning and Data Protection.* (2017) London:ICO https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf
- Jerven, Morten. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*. (2013) Ithica:Cornell University Press

International Council on Archives
Conseil International des Archives

References and Further Reading

- LeSueur, Andrew. 'Robot Government: Automated Decision-Making and its Implications for Parliament' [Draft chapter for publication in *Parliament: Legislation and Accountability* (Oxford:Hart Publishing) 2016] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2668201
- Lorenzi, Jean-Hervé et Mickaël Berrebi. *L'avenir de notre liberté* (2017) Paris : Eyrolle
- Lynch, Clifford. Stewardship in the 'Age of Algorithms' *First Monday* Vol 22 (12) December 2017 http://firstmonday.org/article/view/8097/6583
- Mason, Susan . 'Analysing Stereotypes Across Time Using Google Ngram Viewer' *SAGE Research Methods Cases Part 2* (2018) doi:10.4135/9781526436245
- Mason, S. E., C.V. Kuntz, & , C. M. McGill. 'Oldsters and ngrams: Age stereotypes across time'. *Psychological Reports: Sociocultural Issues in Psychology*, (2015),116, 324–329. doi:http://dx.doi.org/10.2466/17.10.PRO.116k17w6
- O'Neill, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (2016) New York: Crown Publishing
- Venkatesan M *Artificial Intelligence vs. Machine Learning vs. Deep Learning* (7 May 2018) https://www.datasciencecentral.com/profiles/blogs/artificial-intelligence-vs-machine-learning-vs-deep-learning
- Villani, Cédrique. *Donné un sens à l'intelligence artificielle: Pour une stratégie nationale et européenne* (8 mars 2018) https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf
- World Wide Web Foundation. 'Algorithmic Accountability: Applying the Concept to Different Country Contexts'. *A Smart Web for a More Equal Future* (2017) https://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf
- Zambonelli, Franco, Flora Salim, Seng W. Loke, Wolfgang De Meuter and Salil Kanhere. 'The Algorithmic Governance in Smart Cities: The Conundrum and the Potential of Pervasive Computing Solutions' *IEEE Technology and Society Magazine* (June 2018) pp 80-87

Thank you.

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

ICA
International Council on Archives
Conseil International des Archives