Open data and information: Opportunities and challenges for the records profession

Thank you for the generous introductions and the invitation to give this year's Annual Archives Lecture in the Dept. of Information Science at UNISA. It is an honour to be here in a country that has contributed so much to the debate and development of thinking that has informed the archives and records profession, just days after watching your rugby team win the play-off match so convincingly to become the third best team in the world – congratulations!

This evening I would like to consider some of the opportunities and challenges that the open data and information movement presents for us – the role that we records professionals can play in bringing our principles and practice to bear; the new knowledge and partnerships that we will need to develop in order to contribute; and the potential difference we might make.

Slide 2 - Introduction

To do that I will begin by considering the concept of open data and information, its ideology, motivations and aspirations; then I will move on to examine *some* of the key challenges that open data presents in the digital world, and consider the role of the records professional in this space. This will lead me to highlighting important opportunities for our profession and what we will need to be if we are to grasp them and make a difference.

I apologise in advance that, although I know 'data' to be plural, I will use it (strictly incorrectly) in the singular, simply because in spoken language it sounds so much better.

Slide 3 – Open data & information

This is a photograph of an ancient structure less than 20 miles from where I live and work. Hadrian's Wall was built (AD120-130) to establish a 'closed' border between Roman Britain and Scotland, separating civilised Romans from the barbarians! 73 miles long (80 Roman miles), it stretches almost the full width of England, from Carlisle in the west to Newcastle in the east, ending at Wallsend (or Segedunum in Latin). Of course it has been 'open' for centuries, in fact it is a UNESCO World Heritage site, and is now not only a symbol of the past but a source of data and information about our history. I'll use more photographs of HW to **illustrate** my **key messages** this evening.

Slide 4 – concepts of OD

In 2012, UK government Minister Francis Maude (then Minister for the Cabinet Office & Paymaster General) described data as "*the 21st century's new raw material*" in his foreword to the government's *Open data white paper.*[1] In the same document open data is defined as data meeting three criteria. It is:

1. "accessible (ideally via the internet) at no more than the cost of reproduction, and without limitations based on user identify or intent;
2. in a digital, machine readable format for interoperation with other data; and
3. free of restriction on use or redistribution in its licensing conditions." (p8).

More succinctly stated by the Open Data Institute, open data is "*information made available for anyone to use, for any purpose, at no cost.*"[2] 'Anyone' I would argue can include not just human beings but another technology/system or business process that needs to use data for a given purpose. Open data can be of many kinds. Some of it will be 'big' data because it posses certain characteristics, all of which conveniently begin with the letter 'v'. High *volume*; high *variety* (structured, unstructured, semi-structured) and high *velocity* (meaning it is rapidly available for raid analysis). It should also have *veracity* (i.e. integrity and trustworthiness) and *value*. I will return to the last two later.

Open data needs a licence to say it is open and to give the conditions under which the data can be used. Without one data cannot be used or reused, linked to or combined with other data for analysis. Common licence conditions include attribution – i.e. crediting the authors of the data – and share-alike – i.e. releasing the data from any reanalysis, combination or linkage as OD. Typical licences are Creative Commons licences and, for public sector data, Open Government licences.

Ideologically, in the current climate of openness, one might think the main purpose of open data is for greater transparency and accountability, part of the right to information, the right to truth, focused on holding governments, organizations and/or individuals to account, and also adding to the information sources available for communicating history. Indeed the UK government's *Open data white paper* highlights transparency first as a central driver. However, in addition to checking that public funds have been spent appropriately and correctly, open data should enable people to challenge actions and suggest improvements, contribute to public reform, promote stakeholder participation in decision-making, and aid innovation.

This leads to the other main driver for the open data agenda - the promise of *economic growth and efficiency.* Sharing data between public bodies *should* enable them to work more efficiently; open data can be used by other people and organisations leading to innovation, new applications, products and services, improved performance. This is about *getting more from the investment in data collection and capture*, using it for other purposes, combining it with other data and thereby gaining new insights and increasing its *value*. Some might say this is the *only real* driver for OD.

These two motivators are juxtaposed – transparency can be seen as 'retrospective' or reactive, whilst economic growth is forward looking and proactive. But together they aspire to supporting *engagement* (of different stakeholders), *efficiency and effectiveness* (doing things right and doing the right things), and *economic growth* (the appropriate use of resources for innovation).

Examples of the aspirations of open data include combining weather and/or environmental information with information about flora and fauna, or environmental and social

information with health or disease, for a greater understanding, respectively, of the impact of climate on habitats and the effect of the environment on health; enabling researchers, governments and others to address the issues and to plan according to trends.

In a *research* context, open research data is now a requirement of many research funding bodies e.g. the US National Science Foundation, the Bill & Melinda Gates Foundation, the World Bank[3] and the UK research councils because:

> "*Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property."*[4]

So, in addition to requiring open access to published outputs they all require data to be 'archived' and made available. And the purposes are the same - research data provides evidence that the research was conducted properly (transparency & accountability) – and there are well known examples of fraudulent or misleading scientific publications, and data for reuse (secondary analysis) and the generation of further findings and outputs, i.e. "standing on the shoulders of giants."[5]

Just last week the world's first *big data* centre for food and farming was announced[6]. AgriMetrics, based in the UK north of London, will use data and information from the whole supply chain to develop models for improving the *efficiency*, *sustainability* and *profitability* of agriculture. It is about precision farming and wiser use of resources – clean, green farming by developing the metrics of sustainability. Taking information from farmers all over the world, the hoped for benefits are new innovations, products and technologies for farmers alongside policies and international standards in data. Some of the outputs will be OD.

Here in South Africa, Open Data for South Africa[7], operated by the African Development Bank Group (AfDB), provides statistical data on a wide range of topics – from country demographics to water - sourced from the Bank and other international organisations.

There are countless other examples across the world of open data and information initiatives; and it is clear that technology and the internet have been the catalyst for the 21st century open data and information movement. As Prof Nigel Shadbolt said at last year's UK Open Data Institute summit (this year's took place this week) data on the web is the next and much more powerful stage of the Web.[8]

Slide 5 - Challenges

Turning to the challenges, there are a number of dimensions of open data & information that present a range of people, process and technology challenges. These include:

availability - by definition open data is meant to be available. Archives have several decades of experience of widening access to their collections through digitisation programmes. But, despite today's digital world, not everyone has the technology needed to access open data and information, or perhaps the knowledge and skills to use it.

usability - can the data be interpreted, understood and utilised through time? If so, then the *sustainability* of open data collections is an issue. Part of this challenge is *preservation* - the domain of records professionals. Despite the earlier fears of a digital dark age no longer being a concern according to some, because it is tractable[9], sustainability is an issue if only from a resource perspective. Currently open data repositories are often accessible via portals scattered around the Web. Who will maintain the infrastructure, systems and the tools for using the data? What investment is needed, and more pertinently, what is the return-on-investment and for whom? In the UK a number of research repositories have lost central government funding and, in the context of the funding body requirements I referred to earlier, research institutions such as universities are faced with acquiring their own data repositories. JISC recently announced their intention to provide a shared RDM service over the next 18 months, based on demands from the UK academic community, enabling researchers to easily deposit data for publication, discovery, safe storage and long term archiving[10]. Good news.

But I want to focus on the challenges posed by other dimensions, viz. *trust* in and the *quality* of OD; its *value*; *ethical* issues; and *making sense of it all*.

Slide 6 - Trust and quality

First, the issue of *trust*. Trust in what and whom; trust in the quality of the information/data *and* in the provider giving all of the relevant information; trust that the information will only be used in ways that were agreed or consented to, which is particularly important in the context of personal data and research data. Trust overlaps with other dimensions.

Earlier today, at a meeting of InterPARES Trust Team Africa members, I shared details of a project undertaken by one of my colleagues in which trust emerged as *the* key issue. The project analysed the discourse around the UK National Health Service (NHS) *care.data* programme to collect and link together data from all health and social care settings[11] (hospitals & communities) in order to plan and monitor services. A range of care data sets has been collected for a number of years (e.g visits to hospital) and patients are pseudoanonymised using a custom-designed patient ID; but this programme adds data sets from general practices (GPs i.e. local doctors) with the aim of linking GP data with hospital data. Again, though GPs had been providing aggregated data previously, the care.data programme will extract identifiable personal and sensitive data e.g. NHS number, date of birth and coded clinical information. The GP data is far more individually identifiable than hospital visit data, e.g. date of birth compared with age group. It effectively covers the whole population whereas hospital activity covers a small proportion of the population at any one time and is episodic.

Although the *care.data* programme needs to be seen in the wider political context a number of factors contributed to lack of trust including: poor governance, data security, its purpose and, with that, comes informed consent. Purpose lay at the heart of the trust issue with people being deeply concerned about what the data is, what it is being used for and who can access it. If open data is to be trusted then we need to trust both the data and the provider.

Data must be *authoritative* and therefore display some important characteristics. It must be *authentic* i.e. proven to be what it claims to be. It must have *integrity* to confirm its completeness, that has not been altered or that it is clear how it has been altered. And it must be *reliable* i.e. dependable, full and an accurate representation of the activity.

Of course you will recognize these as the characteristics of records, defined in ISO 15489[12], which document and provide proof of a transaction and are information for an organization or person. As records professionals we develop systems and procedures to ensure records have these characteristics and can be trusted. Systems that are reliable and comprehensive, have integrity, be compliant with requirements and ensure records are created, maintained and managed systematically. McDonald and Leveille[13] show how systems design and operation can facilitate data integrity using a fictitious public sector organisation and a pipeline licensing system.

Anne Thurston, well known in sub-Saharan Africa for her work with the IRMT, has highlighted trustworthiness of data as being linked to the trustworthiness of records, as she sees data being derived from records. I agree with her but I also believe that, in addition to records being the source of data, data can also lead to the creation in records. They are inextricably linked and their quality is mutually beneficial. The InterPARES Trust project is working on trust in the context of records (and hence data) on the internet/in the cloud. As I said, Team Africa met today to discuss the interesting projects they are undertaking and new ones proposed.

Trust in the data provider is much more complex and subjective. Whilst good 'corporate' governance and information governance, including systems for data capture, management and security, are crucial, it is often the perceived reputation and track record of the provider that determine their trustworthiness. For example, the BBC has a developed a global reputation as a trustworthy news broadcaster, whereas governments and political parties are often less trusted. Trust in organisations is hard gained yet easily lost, witness the recent example of Volkswagon and the falsification of car exhaust emissions performance and the loss of data from the Internet Service Provider TalkTalk through a cyber attack on their website. Although these examples are not directly related to open data they illustrate reasons for trusting or distrusting potential data providers. On a positive note, the data broker Resultsmarks states that it wants to share reliable data[14].

Trusted digital repositories combine trust in both the data and the data provider and long term access to the data. They are often those based in national archives and hence familiar territory for records professionals.

Linked to trust is the issue of data quality. We can make a judgment of the quality of this part of Hadrian's Wall simply by observing it and comparing it with adjacent parts. We may not question its quality if we find the photo on, for example, the National Trust website as it is one of two organisations that act as its guardians. But do the missing pieces suggest a reduction in quality or integrity? We would only be able to make a judgement based on other knowledge, about the way Romans built walls for example. With data it is more

difficult – is it accurate, is it complete, how reliable is it? The missing elements are not always obvious.

Anne Thurston[15] noted that the release of open data has occurred *"without a methodology for ensuring their accuracy and traceability to reliable information sources. Government data relies heavily on evidence derived from official government records, and in many countries, public records are not managed in relation to international standards.*" If records are not well managed then the data from which they are derived will likely be incomplete and/or inaccurate. I am reminded of the phrase in the early days of computing – garbage in, garbage out. If our RK systems are poor then any data derived from the records will be poor or at least lack quality. The implications are clear – at best mis-information and mis-understanding, at worst inappropriate decision making and action. In a World Bank *Transparency and Information Management Open Discussion Forum* in Oct 2014[16] Jean-Louis Sarbib, CEO at Development Gateway[17], said he saw "wrong information" about the treatment for HIV Aids spread and reach the "highest levels in South Africa" when he was in charge of that programme at the World Bank, with what he described as "devastating results."[18]

However, an open data user may be unaware of potential quality issues, trusting in the data provider, resulting in a danger that they take it at face value and use it in good faith. I am likely to trust data from the UK's Office of Public Sector Information (OPSI) because it is part of The National Archives in whom I have faith from a RK perspective. They set standards for public RK and public sector information, but I do not know the degree to which they audit data that is opened for accuracy and quality They may not be able to and rely on their faith in government bodies to have followed the standards and produced trustworthy, quality data and records.

The trustworthiness of open data begins with system design and data/records creation. Olav Sataslatten (National Archive of Norway) concluded that *"the formative stage of data creation has to be addressed*" if we are to create relevant open data solutions for citizens in the world of 'apps', when he wrote about the importance of metadata from EDRM systems for access to open government data.[19] Thurston has called for "collaborative research on the risks involved in releasing untraceable data… followed by the development of good practice methodology for testing data integrity and accuracy"[20] and John McDonald, formerly of LAC, offers an approach to integrating the RTI into business processes in the context of Open Government and accountability. I think it's an approach that could work more broadly for OD.[21]

Interestingly, the Open Definition (from Open Knowledge) is the only definition I have found that includes the word provenance saying: "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)."[22] The notion of provenance – vital for trust and quality – seems to have been forgotten.

Slide 7 - Value

Value. Figures suggest the data/information we create and copy is doubling in size every two years and will reach 44 zetabytes (44 trillion gigabytes) by 2020[23]. Some of that volume is or can be a valuable asset; for example, the volumes of environmental and climate records that show trends in weather patterns and help to predict weather events, or population data that helps us plan how to feed people; health records that enable us to track and fight disease etc. But is it necessary to retain all information that is created forever, even for long periods, despite the fact that it seems easy to do so in the digital world? I would argue not since, according to IDC, the vast majority of information, typically well over 90%, is not archived for permanent retention nor used more than once.[24] It would be irresponsible, leaving the problem to the next person/generation.

The volume is a digital iceberg, the tip being OD; but who shapes the tip of the iceberg? Who determines what we should make open and how? How do we decide, from this unprecedented volume of information, what is valuable to share and make open?

At the World Bank *Discussion Forum* I just referred to Anders Pedersen, an independent journalist, and Jean-Louis Sarbib made it clear that the focus had been on the supply side, often selecting data that was easy to open, rather than on the basis of its value or demand. Sarbib cited the example of the Kenya OD[25] portal which had attracted a lot of interest initially amidst much publicity, but whose use had not been sustained, with access falling off dramatically after just two months. He called for more focus on the demand side for OD.

Whilst I acknowledge that there are different perspectives on the impartiality of the role of the records professionals we do have expertise and principles for identifying what is (or potentially is) important or valuable, and ensuring the data and records are available and interpretable only for as long as required. This is appraisal - identifying what records should be created, kept and for how long; separating the wheat from the chaff and deleting what is no longer useful. Decisions are based on an assessment ofbusiness functions and processes in the context of the regulatory environment, business and accountability requirements, risks and future needs of internal and external stakeholders. Open data needs appraisal.

Archivists understand their users and, if we understand more through information behaviour and log analysis etc (as librarians have done) then we should have something to say about the potential demand for and selection of data to make open. Making everything open is not possible and opening data that has no or limited demand is not economically viable. It negates one of the main purposes of the open data movement.

Slide 8 - Ethics

The motivations for open data are all positive but there are some important ethical issues to consider and there can be *unanticipated consequences.* From an open data perspective, it is so much easier to search and find digital information and to make links and discover connections that would have been highly unlikely, if not practically impossible, in the analogue world. In Australia, for example, the National Library's Trove online service[26] includes the ability to full text search the now digitised Australian newspapers, the analogue content of which is already in the public domain. One of the site's FAQs is '*could you remove an article containing personal or family information?*' The NLA's response is "*We appreciate*

that some people are finding surprising information about themselves or their relatives which is sometimes good and sometimes bad, and that this may be of concern" but their disclaimer clarifies that they do not review or censor the newspaper articles[27].

An interesting education example is given in the 2014 report on '*Big data and privacy: a technological perspective*' from the US President's Council of Advisors on Science and Technology. Access to the log information of online courses, including MOOCs, will make it possible to create and maintain longitudinal data about learner engagement with learning materials and activities; whether they repeat or skip content, their attention span etc.. When linked to grades this will help improve education - but, if these are tracked over time and linked to an individual's future success, then there are significant privacy issues. The report notes that:

"[k]nowledge of early performance can create implicit biases that color later instruction and counseling. There is great potential for misuse, ostensibly for the social good, in the massive ability to direct students into high- or low-potential tracks."[28] (i.e career paths)

Yet it is precisely the linking of different data(sets) – mash-ups - that brings the opportunities for new knowledge or business. Safe havens (trusted third party organisations) have been proposed where identifiable data is held securely and - on demand - can be linked with other data. The researchers are then given only the anonymised, aggregated results of this linkage.

However, anonymisation is a complex activity, which is much more than the removal of participants' names. There is a problem that for very sensitive qualitative data, anonymisation may strip the data of too much content so that it becomes impossible to reanalyse it. People can also be indirectly identified by linking up separate pieces of information from different sources; each individual piece of information may not identify a person but when linked together identification becomes possible (the average number of data points required is only four). As Richards and King noted *"privacy protections focused on personally identifying information are not enough when secondary uses of big data can reverse engineer past, present and even future breaches of privacy, confidentiality and identity*."[29] This has become a much easier operation because of the amount of data that exists on the internet about every individual - our personal trails in the e-world.

Ethical concerns about making data obtained from human participants open on a long term basis are not completely addressed by consent and anonymisation. Open data in digital form made available via the internet to a potentially global audience can be readily copied and kept indefinitely. Sophisticated tools open up new ways to analyse and manipulate the data and make data linkages as I've already said. Therefore, when data is shared we cannot anticipate who will use it, why and how. For example, in 2011 the cigarette manufacturer, Philip Morris International used an FOI request to obtain research data about teenagers smoking habits from researchers at the University of Stirling.

In an open research data environment, informed consent is more complex. Ethically we cannot ask participants to give blanket consent to data sharing. The informing part of the consent requires them to fully understand concepts and practices of data sharing that may

be unfamiliar to them. Consent may need to be much more nuanced and sophisticated, with participants restricting access for certain purposes, certain types of researchers etc.

Balancing privacy, confidentiality and security with access, sharing and re-use in the world of open data is a complex and emotive area. There are tensions between the right to privacy, confidentiality and security (data protection) and the right to information and sharing for the benefits I have already highlighted. Sometimes these conflict. Earlier this year Facebook announced it was adding *more* features to enable users who see posts from their friends that are concerning to them (to report them so that Facebook could get involved. Concerns might relate to self-harm or suicidal thoughts. Facebook will look into the post and if they think the person is in danger of harming themselves, they will contact them with a message to talk with someone – the messages are fairly basic. Although Facebook worked with relevant organisations (eg suicide prevention) there were negative reactions to using data in this way.[30]

Slide 9 - Sense-making and decision making

The final challenge I want to raise is perhaps the greatest one – sense-making.

I recently heard the Vice President of Intel (Genevieve Bell) suggest that the Domesday Book was the first example of big dataset in the world.[31] In 1086 William the Conquerer wanted to know what his kingdom comprised, in terms of the citizens, those in charge, the landscape etc. He wanted to *make sense of his world*. A primitive database true, but one carried by the King and consulted on his travels for a range of purposes. Over 900 hundred years later its contents are open.

Today the purpose (value) of open data is to free people "to make informed choices" about how they live, what they buy, who they vote for. A world where, the Open Knowledge Foundation states "information insights are accessible – and apparent – to everyone."[32] Ideologically I cannot argue with this but what is the reality? Making sense of data is challenging.

Yes, we have increasingly powerful software and new data analysis tools to better understand consumers and customers, science, nature or society, or for intelligence. But are we drawing the right conclusions, making appropriate decisions? Do users have the capacity and capability to do so?

What does this photograph show? A tree in a hollow. But where is it, what kind is it, how old is it? Is there anything special or significant about it? In the context of this evening's lecture, and given what I said earlier about the photographs, you might infer that it is somewhere near Hadrian's Wall. You would be correct – in fact if you look closely you can see the wall running up and down the hills to either side. I could drive to it (or search the web!) and give its precise GPS location. Though I don't know how old it is, I do know that it is a sycamore tree and its location is known locally (and beyond) as sycamore gap. What is special or significant about it is that it features in the 1991 film *Robin Hood – Prince of Thieves* starring the American actor Kevin Costner. But without this and much more contextual metadata, we cannot make full sense of it.

Sense-making requires *sufficient context* in order to understand the data and information being accessed and is very challenging in the digital environment. From an open data perspective, knowing where information came from, in what circumstances it was created, who authorized or approved it, and its relationship to other information are vital not only for establishing its authority but also for fully understanding and making-sense of it. Again Jean-Louis Sarbib provides a good example in the World Bank forum about aid data from China. Because China was not publishing it the data was gathered from news announcements etc and published by the World Bank with a disclaimer about its source. During the discussion a participant made exactly this point about the value of having that contextual information/metadata associated with the data set to attest to its quality.

From a records perspective context "includes information about the business processes in which the records are created" allowing users to understand the reliability of the records creator, the environment in which records were created, the purpose or business activity being undertaken and their relationships with other records or aggregations. It also includes information about the systems in which they are managed, the organisations that manage them, and their broader operating context. Sufficient information about these different layers is needed to make the records understandable and therefore useable to users (ISO 23081-1) and, since context may change, this information will accrue through time. The same is true of data.

In the analogue world, context was often apparent by looking at the record, or the file of which it is a part. Its form might signal its formality; the letter head would give the details of the organisation, the file an indication of the business process or activity. In the digital world this is often not clear.

In the digital open data world we must determine what metadata should be created with the data (or record) and through its processing, how that metadata will be persistently linked and managed. But contextual metadata is very challenging because its creation and capture can be time consuming, if not expedited automatically through careful systems specification and design. Records professionals have an important role to play here, yet I am not sure that is happening enough. Contrast that with those in arts, culture and communications, where researchers at the University of Copenhagen are looking at 'correct' interpretation of big (open) data from a humanistic angle.[33]

Slide 10 - Opportunities

Turning to the opportunities for records professionals, in discussing these challenges I have already highlighted some – ensuring the quality and trustworthiness of OD, that ethical issues (as well as legal/regulatory ones) are considered and addressed, that we identify what is valuable and in demand rather than supplying what might be or is easy, and perhaps most important of all, enabling users to make sense of data so that they can make appropriate decisions and take appropriate actions. Of course it is not the sole responsibility of the records professional to do this, we are part of the wider open data space inhabited by technologists and computer scientists, statisticians and mathematicians, senior executives, lawyers, data creators and users etc. But these opportunities are about bringing our

principles, practice and understanding of who creates information and how it is created, managed and used, to the open data table.

However, how many of us can claim to be at the table?

How many people here this evening are or have been involved in open data initiatives? How many would consider they have a role in the open data and information arena? How many would say you have expertise to offer?[34]

The open information agenda is not new for records professionals. Archivists have always provided access, sometimes controlled, to their collections - for history, research, innovation, social and personal interest. Records managers in the public sector proactively and reactively respond to information access requests. However, there is little evidence in the published literature that we are discussing it or making it clear to others that we have an important role to play, making our voices heard. As editor of the *Records Management Journal* I first drew the opportunities in this space to the attention of the journal's readers in a 2012 editorial[35], the same year that (independently) Anne Thurston wrote about trustworthy records and OD. This led to what I believe to be the first issue of any journal in our field being devoted to open data and big data, which I invited Anne to guest edit.[36]

I see two important opportunities for records professionals - leadership and education and training.

First, in addition to being collaborators, records professionals must demonstrate leadership in the open data arena – see the near and far horizon - else others will, particularly IT professionals. Only last week the White House released the third US Open Government National Action Plan, which the National Archives were involved in developing, and US Archivist (David Ferrerio) blogged about the work NARA would to lead. I and colleagues have also tried to do this.

Prof Michael Moss is working with computer science colleagues on technology assisted methods for reviewing the sensitivity of UK government records prior to public release – unstructured open data of the kind commonly known to archivists. In the context of the unanticipated consequences of search that I spoke of earlier, if appropriate and accurate sensitivity review cannot be assured there is a risk this will lead to precautionary closure of records and data. There are implications for social and historical research and the (potentially more limited) ability of citizens to challenge conclusions and hold government to account – the very antithesis of the open data movement. [37]

Records professionals understand the need to identify potential sensitivities; we have review and redaction processes. These will be needed to enable some data and information to be made open but new tools will be required to do so efficiently whilst maintaining the quality of the review in the face of the digital tsunami. One approach is to use sophisticated information retrieval algorithms that employ techniques such as archival diplomatics to identify potentially sensitive information, by looking for names that might be sensitive or combinations of entities that could identify individuals, such as a name and date of birth, a role, a place etc.[38] Project Abacá, a feasibility project between Glasgow and Northumbria

universities, has developed such algorithms. Still nascent, these and others under development elsewhere, "will be able to rank sensitivity, prioritizing instances of possibly the highest sensitivity."[39] They illustrate leadership and collaboration. They also have potential application in other contexts and sectors, such as open data and e-discovery and in some way respond to Thurston's calls for research into risks.[40]

In relation to the sensitivity review work, in two weeks' time we will be holding a conference called *Threats to Openness in the Digital World* to discuss the issues and identify actions. Amongst a number of leading speakers will be Sir Alex Allan, author of the report for the UK government on a review of the annual release of government records.[41] This is another example of leadership and also education, which is our second important opportunity.

An exemplary demonstration of leadership can be found at Girona City Council where the records management department lead the location of datasets and selection of data to open using records management principles and tools. Why? To "be useful to the organization … and to reinforce its position in the current trending to data management."[42]

Slide 11 - Opportunities

We need to consider open data challenges in designing our education programmes, but educating the relatively small numbers of records professionals in the world is not sufficient.

Our current work on education is the development of a bid, with five other European iSchools, for an Innovative Training Network for new data professionals in the context of information governance, open data and big data. As someone from a leading publisher suggested to a colleague involved in the bid: "*in five years, analyzing data will be a commodity. Relating data chunks and getting meaning out will be the tricky part*". Hence our view of the future data professional is someone who has a 360^{o} perspective and understanding. Perhaps not an expert in every aspect of open data but very knowledgeable about all aspects and someone who can communicate and collaborate with others to make it happen. It is much more than technology and tools expertise, it concerns the entire data value chain, processes, creators and consumers.

Just as the National Trust and English Heritage educate visitors to Hadrian's Wall about the underfloor heating system they developed at Housestead Fort (picture in this photograph), we also have a very important role in educating open data consumers and creators so that they can create and make available good quality, trustworthy, data for others to make sense of, making appropriate decisions and taking appropriate actions. Here our work to develop a research data management skills training programme[43]. By leading and developing new collaborations/partnerships (other university staff and services Grad School/Library and leading UK digital curation/ preservation bodies the Digital Curation Centre and Digital Preservation Coalition) the DATUM projects are enhancing the knowledge and skills of PhD students and staff in managing their research data and to improve RDM in practice. This went beyond records management advice and guidance, covering some of the open data challenges I raised earlier - concepts of anonymisation and consent.

Of course we must not forget our own professional development and education; that is essential if we are to rise to the challenges because of their complexity and multi-disciplinary nature. We too need to be innovators and be able to utilise technology in effective and imaginative ways, which depends on our own level of 'digital literacy'. Are we up to it? In a podcast for students on a records and IG module I taught earlier this year, Paul Mullon, a fellow member of the ISO standards RM committee, questioned whether records professionals are equipped for the challenges facing them today. At least part of the responsibility for this lies with people like myself – educators – but part of it he said related to the personalities and interests of people who enter the profession, liking archives for what they contain. He asked whether the challenges today are more suited to a different type of person. So we not only need to ensure we offer the right knowledge and skills but also attract a wider range of people into our profession, and collaborate with a wider range of people with complementary expertise.

Slides 12 & 13 - Conclusion and the Paradox of our Age

To conclude, open data is exciting and, when done well, offers huge opportunities for the economy, combating health, other social and environmental problems, increasing transparency and improving accountability. But with it come dangers of the kind I have mentioned. Records management principles, such as the characteristics of authoritative records, the design of good RK systems, metadata, appraisal and retention management, can and do support access to trustworthy, quality open data and information that is ethically available, is valuable and available in ways that enable sense-making to underpin appropriate, reliable decision making, recommendations and action. Undoubtedly there are some non-trivial challenges a selection of which I have only been able to scratch the surface of here. These challenges present opportunities as I have tried to illustrate; ones that require new approaches, technology assistance, and mostly importantly leadership and education.

Three years ago I discovered a verse on a book mark made by Tibetan refugees in India and entitled '*The Paradox of our Age'.* Attributed to the current Dalai Lama, though disputed by some, the words share a clear message about what we have and do, what we have lost and do not do. Some of those words are particularly pertinent for the challenges and opportunities for records professionals in the context of open data and information. The book mark I bought reads thus:

We have bigger houses but smaller families;
more conveniences, but less time;
We have more degrees, but less sense;
more knowledge, but less judgement;
more experts, but more problems;
more medicines, but less healthiness;
We've been all the way to the moon and back,
but have trouble crossing the street to meet
the new neighbor.
We build more computers to hold more
information to produce more copies than ever,

but have less communication;
We have become long on quantity,
but short on quality.
These are times of fast foods
but slow digestion;
Tall men but short character;
Steep profits but shallow relationships.
It's a time when there is much in the window,
but nothing in the room.

I have emphasised the parts I feel are particularly relevant to us as records managers and archivists whose focus is on the creation, capture, organization and management of the increasing quantities of digital information, whose value lies in its appropriate communication and use, to make decisions or judgements. The words may help you reflect, as I did, on how well we are succeeding.

I believe records professionals can make a very important contribution to better OD. However, I challenge our profession – and that includes myself – to recognize the opportunities, understand the requirements and rise up to the challenges so that we can become long on quality, shorter on quantity and ensure there is much in the open data room that will make a difference.

Slide 14 - Acknowledgements

I will end by acknowledging the open source photographs I have used this evening, and leave you with a final one. Another spectacular open space in Northumberland - Embleton Bay looking at the ruins of another historic construction, Dunstanburgh Castle.

Final slide

Thank you.

Slides

3. Open data & information: NY7266 Hadrian's Wall and Turret 41a. 3 km from Whiteside, Northumberland, Great Britain http://www.geograph.org.uk/photo/846921
4. Concepts : Housesteads Fort, Hadrian's Wall, Northumberland, Great Britain https://www.flickr.com/photos/littlemisspurps/5479985746/
5. Challenges : NY7567 Hadrian's Wall, Peel Crags. 3 km from Henshaw, Northumberland, Great Britain http://www.geograph.org.uk/photo/3189381
6. Trust & Quality : https://www.flickr.com/photos/michaelloudon/5848861168
7. Value : Hadrian's Wall, Milecastle 39 https://en.wikipedia.org/wiki/File:Milecastle_39_on_Hadrian%27s_Wall.jpg
8. Ethics : Housesteads Fort, Hadrian's Wall https://www.flickr.com/photos/68259253@N00/474320259
9. Sense-making : Sycamore Gap, Hadrian's Wall, https://www.flickr.com/photos/michaelloudon/5848110111/
10. Opportunities : NY7467 : Hadrians Wall 4 km from Melkridge, Northumberland, Great Britain http://www.geograph.org.uk/photo/946902

11. Opportunities : Underfloor heating at Housesteads Fort, Hadrian's Wall, Northumberland, Great Britain https://www.flickr.com/photos/quinet/13303511835/
12. /13. Conclusions : Lone tree near Caw Gap Sill, Hadrian's Wall, Northumberland, Great Britain. Anita Nicholson https://www.pinterest.com/pin/522065781776821546
14. Final slide : Dunstanburgh Castle looking south from Embleton Bay, Northumberland, Great Britain https://www.flickr.com/photos/96018577@n00/9676296596

References

[1] *Open data White Paper: Unleashing the potential*. *CM8353*. HM Government, 2012. www.official-documents.gov.uk

[2] The Open Data Institute http://theodi.org/. Established in 2012 by Sir Tim Berners-Lee and Professor Nigel Shadbolt to increase awareness of open data and sharing principles and practice.

[3] Clobridge, Abby. (2015). Open data: Shining a light on data management practices. *Online Searcher*, V39 (4), page: 68-70.

[4] Research Councils UK. *RCUK common principles on data policy.* http://www.rcuk.ac.uk/research/datapolicy/

[5] "If I have seen further, it is by standing upon the shoulders of giants." Attributed to Bernard of Chartres, 12th century and Sir Isaac Newton, 1676.

[6] BBC. *Farming Today*, BBC Radio4, 26/10/2015. George Freeman, Conservative MP for Mid Norfolk, talking to Charlotte Smith about Agrimetrics, the first big data centre for food and farming, established by Innovate UK as part of a network of Agricultural Innovation Centres. http://www.bbc.co.uk/programmes/b06kb7z1 Agrimetrics website: http://www.agrimetrics.co.uk/

[7] Open data for South Africa http://southafrica.opendataforafrica.org/

[8] Shadbolt in a video clip about the 2014 ODI summit cited open data on the web as the next stage of the web's development http://summit.theodi.org/2014-summit/

[9] Kilbride, W. (2011). Aiming for obsolescence. (Editorial). *What's New*, April. Digital Preservation Coalition. http://www.dpconline.org/newsroom/whats-new/684-whats-new-issue-35-april-2011#Editorial35 Milic-Frayling, N. (2014). Sustainable computation as a means for digital preservation. *2nd Annual Conference of the ICA, Girona, October 2014*. Pre-paper notes in: Evaluation and strategies of digital preservation and UNESCO's role in facing the technical challenges van Gorsel, M., Leenaars, M., Milic-Frayling, N. and Palm, J. http://www.girona.cat/web/ica2014/ponents/textos/id100.pdf

[10] JISC RDM Shared Service Pilot. http://researchdata.jiscinvolve.org/wp/2015/10/07/jisc-rdm-shared-service-pilot/

[11] Childs, S. and McLeod, J. (2015). *A case example of public trust in online records: the UK care.data programme. Interim report of a project undertaken for the InterPARES Trust project*. https://interparestrust.org/

[12] ISO 15489-1 (2001). *Information and documentation. Records management. Part 1: General*. ISO.

[13] McDonald, J. and Leveille, V. (2014). Whither the retention schedule in the era of big data and open data. *Records Management Journal*, V24(2), pp. 99-121.

[14] US Environmental Protection Agency (EPA). (2015). EPA, California notify Volkswagen of Clean Air Act Violations, Carmaker allegedly used software that circumvents emissions testing for certain air pollutants. Release Date: 09/18/2015 http://yosemite.epa.gov/opa/admpress.nsf/6424ac1caa800aab85257359003f5337/dfc8e33b5ab162b985257ec40057813b!OpenDocument ; TalkTalk website cyber attack on 21/10/2015. See http://help2.talktalk.co.uk/oct22incident; ResultsMark https://www.resultsmark.org/home

[15] Thurston, A. (2012). Trustworthy records and OD. *J. Community Informatics*, V8 (2).

[16] *Transparency and Information Management Open Discussion Forum 22 Oct 2014*. Presented by Anders Pedersen and Jean-Louis Sarbib. Chaired by Dr V Lemieux. https://www.kaltura.com/index.php/extwidget/preview/partner_id/619672/uiconf_id/26354192/entry_id/1_ua5j0ab7/embed/auto

[17] Development Gateway, an "international non-profit social enterprise creating information solutions and cultivating skills to turn data into lasting results" http://www.developmentgateway.org/

[18] Sarbib speaking at *Transparency and Information Management Open Discussion Forum 20141022*. Op. cit.

[19] Sataslatten, O. (2014). The Norwegian Noark modeal requirements for EDRMS in the context of open government and access to governmental information. *Records Management Journal*, V24(3), pp.189-204.

[20] Ibid p5
[21] McDonald, J. (2015). Integrating open government and RTI [right to information] in business processes for accountability. *Transparency and Information Management Open Discussion Forum 7 Jul 2015.* Presented by John McDonald https://www.kaltura.com/index.php/extwidget/preview/partner_id/619672/uiconf_id/29215882/entry_id/1_elg5o1ve/embed/auto
[22] Open Definition. http://opendefinition.org/
[23] 44 zettabytes is 44 trillion gigabytes. See: IDC. *The digital universe of opportunities: rich data and the Increasing value of the Internet of Things*. April 2014. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
[24] Ibid.
[25] Kenya open data Portal https://opendata.go.ke/
[26] National Library of Australia. Trove http://trove.nla.gov.au/
[27] National Library of Australia. Trove FAQs http://trove.nla.gov.au/general/using-digitised-newspapers-faq/ and disclaimer "content which was published legally is not censored" http://trove.nla.gov.au/general/about#disclaimer
[28] Executive Office of the President. President's Council of Advisors on Science and Technology. (2014). *Big data and privacy: A technological perspective*. US Government. https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf (p14)
[29] Richards, N. and King, J. (2014). Big data ethics. *Wake Forest Law Review*, V49, p.393-432. http://pacscenter.stanford.edu/sites/all/files/RIchards%20and%20King%20Ethics.pdf (p393)
[30] Suicidal Thoughts & Tendencies Prevention: Facebook Adds New Features to Make Reporting Suicide, Self-Harm Related Posts Easier. Robert C. Weich III (staff@latinpost.com) First Posted: Feb 26, 201
[31] BBC Radio 4. *Peter Snow returns to the future*. 22 Oct 2015. Guest Genevieve Bell. http://www.bbc.co.uk/programmes/b06j5qcz
[32] Open Knowledge Foundation https://okfn.org
[33] Larsen, RW. (2014). Researchers to reveal the dangers of 'Big Data'. *Science Nordic,* 21 Nov 2014. http://sciencenordic.com/researchers-reveal-dangers-%E2%80%98big-data%E2%80%99. Original story in Danish at http://videnskab.dk/teknologi/forskere-skal-afslore-farerne-ved-big-data
[34] Note: only three people indicated they were/had been involved in open data initiatives, a few more considered they had a role and hardly any hands were raised in answer to the final question about expertise. However, speaking with a participant after the event suggested this may not have been an accurate reflection of reality but a hesitation on the part of the audience.
[35] McLeod, J. (2012).The paradox of our age. *Records Management Journal*, V22 (3), pp.148-151.
[36] Big data, open data special issue, *Records Management Journal*, V24 (2), 2014.
[37] As discussed by Moss, M. (2012). Where have all the files gone? Lost in action points every one? *J. of Contemporary History*, V47(4). pp. 860-875. http://nrl.northumbria.ac.uk/13176/
[38] Moss, M. (2015). What is the same and what is different. In: Moss, M., Endicott-Popovsky, B. and Dupuis, M.J. (Eds). *Is the digital different? How information creation, capture, preservation and discovery are being transformed*. Facet Publishing.
[39] Ibid.
[40] Thurston, *Op cit.* p. 5
[41] Allan, Sir Alex. (2014). *Records Review.* 6/11/2014. https://www.gov.uk/government/publications/records-review-by-sir-alex-allan
[42] Casellas Serra, L.E. (2014). The mapping, selecting and opening of data: The records management contribution to the Open Data project in Girona City Council, *Records Management Journal*, Vol. 24 Iss: 2, pp.87 – 98. DOI http://dx.doi.org/10.1108/RMJ-01-2014-0008 (p.88).
[43] Northumbria University. (2012). *DATUM: research data management*. Available at: http://www.northumbria.ac.uk/datum