Traces through Time:

A probabilistic approach to connected archival data

Sonia Ranade
The National Archives
Kew, UK
sonia.ranade@nationalarchives.gsi.gov.uk

***Abstract* — This paper explores challenges in applying modern data analytical techniques to historical archival data, through the lens of The National Archives' *Traces through Time* project. It considers the potential of computational approaches to create new access routes for heritage collections, offers a practitioner's perspective on the implications for archival description in a digital age and anticipates future developments in this field.**

Keywords - archives; description; linked data; big data; probabilistic

I. THE CHANGING NATURE OF ARCHIVES

A. *Archives as data*

Archives and archival data are changing. New accessions to archives are increasingly digital – whether born-digital or digitised for accession [1] – and existing collections are being digitised at an accelerating rate[1]. This is producing archival data of such scale, complexity and coverage that it is opening up new avenues for research, with an associated demand for new computational tools and approaches for researchers and archivists alike. At The National Archives we are witnessing a transformation in the nature of our collections: from the archive as static boxes of documents, to the archive as fluid, conceptually interconnected data.

This influx of digital content is disrupting our long-standing approach to archival description. In parallel with growing volumes of records, we are seeing a shift in the nature of the archival file and a need to manage and provide access at a lower level of granularity: to individual digital objects, instead of papers collected together under a physical cover. It is becoming unsustainable for our archivists to facilitate access to these records through creating individual authoritative, high-quality archival descriptions. And our transferring departments are no more able to do this than we are. Our current descriptive standards [2] do not offer a sufficiently pragmatic or flexible approach to describing the range of material we are receiving and the task of applying current descriptive practice to the 'digital heap' is rapidly becoming a barrier to the transfer and accession of digital records. However, the changes that make our current approach unsustainable also bring their own opportunities. We are moving from a world in which archival data is metadata and the value of the record is locked away in scanned images or paper[2], to one where the whole record can be mined and analysed. Our inability to create fine-grained descriptions need not make us poorer, but we must be ready to seize the opportunities that the shift to digital offers.

B. *The second-generation digital archive*

Digital public records are fundamentally different to their paper equivalents. Every step from their original creation and use to their eventual archiving and preservation has required new thinking, process and technology. However, the access we provide to this material, via a 'first generation' of digital archival catalogues is modelled on access to physical collections. The National Archives' *Discovery*, the first comprehensive online national archival catalogue, is ground-breaking in many respects but still has an underlying access model that is predicated on a human reader viewing one document at a time. It is only with the recent explosion in digital accessions that we are beginning to recognise the potential that digital collections offer for enabling access and research.

Incremental development of on-line catalogues has already brought new research questions within reach, by exposing archival descriptions to tools for indexing, searching and linking. However, The National Archives' emerging digital strategy [3] envisions a second-generation digital archive: a disruptive digital archive that is digital by instinct and design [4]. This second-generation digital archive does not yet have a clear form. Neither do we fully understand the scale or scope of the challenges it will present for computer-science and archival science research. As archivists, we will have the chance to re-imagine our approach to access, and will need to embrace computational thinking if we are to deliver meaningful access to a greater diversity of digital content. From a practical perspective, we anticipate that some of the answers will be found in existing computational techniques adapted to fit this new problem domain, while others will rely on developments that are currently at the forefront of computer science research.

[1] Several significant new collections have been released by The National Archives in recent years. See, for example: www.gov.uk/government/news/thousands-of-war-heroes-wills-released and www.digitalarchives.bt.com/web/arena

[2] 'Paper' is used here as shorthand for the wide range of physical archival media in The National Archives' collections. These include paper, parchment, vellum and photographic material.

C. *Computing over archival data*

Digital archival collections lend themselves to computational approaches to analysis, interpretation and access. But because The National Archives is not yet a natively digital archive, each application of a new technique raises challenges for the way we store, compute over, publish or think about our content. In our work on a recent project, *Traces through Time* (TTT), we encountered a number of these challenges, which give us insight into some of the barriers to computing over our collections.

TTT explored the application of computational and statistical methods to linking archival big data. The project had the combined aims of improving decision-making on records-closure; facilitating access through the creation of new routes for navigation; and creating linked datasets and linking methodologies of research value in their own right.

This paper will explore the challenges we encountered in applying modern computational techniques to historical archival data, through the lens of TTT. It will examine the potential for new tools to transform archival research and consider the role of archives in facilitating new modes of access and use of heritage collections. This paper will give a practitioner's perspective on the possible implications for archival description in a digital age and will attempt to anticipate future developments in this field.

II. TRACES THROUGH TIME: LINKING PEOPLE WITH CONFIDENCE

Traces through Time investigated a crucial question in researching large volumes of data and making robust access decisions about that data: that of identifying and linking individual people in a rigorous way across large, diverse and distributed datasets. This is a first step towards The National Archives' vision of a 'distributed national collection' with Discovery as the primary destination for anyone wanting to access archives in the UK [4]. Solving this would help connect previously isolated sources, linking datasets in new and powerful ways to enable both macro and micro research, whilst accelerating the release of name-rich records into the public domain. Advances in this area therefore have the potential to facilitate genuinely radical shifts in access to archives that could transform the research landscape.

The project described here aimed to begin this transformation by extending the boundaries of current computational archival science research in three important directions: to increase the extent and diversity of the data that can be handled using modern data-analytical techniques; to improve support for the 'fuzzy' data that is typical of archival collections (i.e. data that is incomplete, inaccurate, inconsistent or uncertain); and to develop robust confidence measures for the links we identify, enabling archivists to qualify the assertions we make about records and allowing confidence thresholds to be tailored to fit specific research aims.

The networks of links identified by the project allow researchers to traverse the collections through patterns of connected individuals, bypassing institutional boundaries and hierarchical record structures. Such cross-cutting navigation is by no means unique [5], however it differs from existing approaches in two important respects. Firstly, these links are neither curated nor authoritative. They are automatically generated at scale and are not individually checked. Secondly, they are not bald assertions: every link is qualified with an indication of our confidence in that link. And because we can apply different linking methods at different times to generate multiple links with multiple confidence measures, we must also hold information about the provenance of each link. This work has created data of a size and shape that will not 'fit' into a first-generation digital archival catalogue such as Discovery.

A. *Project approach*

The project's technical approach was one of identifying links through 'fuzzy' comparison of attribute values; evaluating confidence based on statistical techniques and supporting data; and making these links available to researchers. A number of additional processing steps were required to prepare data, configure linking methods and deliver performance at scale. Of these, the data and publication aspects have greatest direct relevance to evolving archival practice and are discussed below.

B. *The TTT 'person' schema*

The concept of the 'person' is at the heart of this project, and we required a data architecture that could potentially encode information about every individual who appears in a public record. There are published schemas for person [6] however, these will neither accommodate the variety of identifying data attributes encountered in our records, nor encode the fuzziness that is a feature of archival data. In early records, an individual may be as clearly characterised by their occupation and relationships to people or places as they are in modern records by their name and date of birth. We wanted to avoid the familiar route of standardising descriptions to create a dataset that is easy to work with, because our understanding of confidence relies on preserving the nuances and uncertainty present in our data. Our TTT person schema [7] gives us a standardised model for capturing this variation. As an example, the extract below illustrates our approach to recording an inferred date of birth where this is not clearly stated in the record.

:age

end	{int}	optional	*(46)*
name	{string}	optional	*(mid-forties)*
start	{int}	mandatory	*(44)*
type	{list}	optional	APPROX|EXACT

*Use age range (***start** *and* **end***) where applicable (otherwise* **start** *only)*
Use **name** *to record the original textual description of age*
Use **type** *to record whether the age is exact or approximate*

Example: *"Harrison was only 14 when he first auditioned for The Quarrymen in March 1958"*
(x:person)-[:hasFamilyNames {order: 1}]->(:familyName {name: "Harrison"})
(x:person)-[:hasAge {dateStart: 195803}]->(:age {start: 14})

Example: *"Harrison celebrated his 21st birthday at the height of the 60s"*
(x:person)-[:hasFamilyNames {order: 1}]->(:familyName {name: "Harrison"})
(x:person)-[:hasAge {dateStart: 1960, dateEnd: 1969, type: "APPROX"}]->(:age {start: 21})

Figure 1. Example node from TTT 'person' schema

C. Data quality

The rates and types of errors we encounter in the records vary significantly between datasets of different periods, collected and prepared in different ways. In this regard, our understanding of our own collections is incomplete. Whilst curation practices place great emphasis on the provenance of our records, we do not generally capture provenance for our metadata. Information such as whether data was created by keying or OCR techniques or whether the original records were typed or handwritten is not generally held within the catalogue, but this knowledge can help us tune our algorithms to improve the quality of linking (Figure 2). Other contextual information also has a bearing on data-quality, for example, some service records may have a high probability of containing falsified dates of birth. Researchers and archivists possess this tacit knowledge but if we are to use it in computation it must be encoded. In capturing this knowledge we are creating new data which qualifies the data we already hold, helping us understand how to work with our records.

This information cannot readily be accommodated within the catalogue or our person schema so, in order to expose it to the linking and confidence algorithms, we have had to create a knowledge base, parallel to but separate from the 'series' level of our catalogue, to record factors that affect the shape and quality of the data.

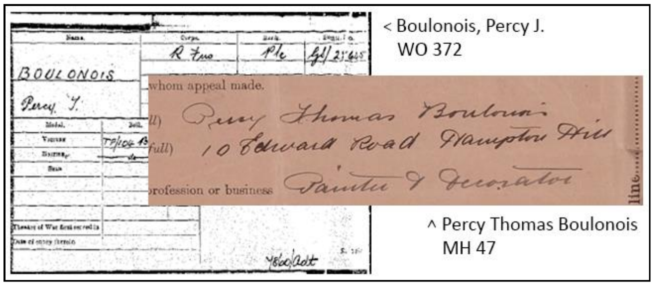

< Boulonois, Percy J.
WO 372
^ Percy Thomas Boulonois
MH 47

Figure 2. Example of a transcription-error 'pattern' from manually keyed early 20th century handwritten documents

D. Probabilistic approaches to inference and confidence

The project takes a probabilistic approach to data linking, in which each link we identify is further qualified with information such as the date the link was asserted, the method(s) used to identify that link and a measure of our confidence that the two linked occurrences of a person refer to the same individual.

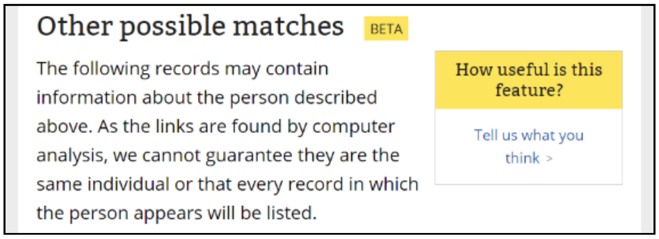
The project deals with other types of assertions that are also probabilistic in nature, such as our confidence in a date of birth or a name in the data itself. However, beyond an indication of whether an attribute value is approximate or exact, our data model does not currently allow us to express our confidence in these assertions.

Inference techniques can be applied to model the values of missing data attributes: a missing date of birth might be modelled by a range of plausible dates of birth. For example, for a Second World War naval service record with no date of birth, we might assume that the individual was aged between 14 and 35. In large collections, we can further extend this approach. An examination of the range and frequency of attribute values across the whole series allows us to generate a probability distribution for the values of that attribute for the cohort in question. Such a distribution bears little resemblance to a 'date of birth' entered in an archival catalogue, but it has real utility for automated processing of the records.

Our longer-term goal is to move beyond the linked pairs of individuals to analyse chains of associated records, and to apply techniques for inferring information from these connections. At that stage our TTT data model will need to be extended or, perhaps, replaced with something new. We are investigating graph-database approaches to managing this data.

E. Publishing and user-experience design

The TTT links are published into Discovery via a new 'Other possible matches' feature. In a departure from our usual practice, these links are a computational, non-curated approach to augmenting archival descriptions. There was some initial concern from within The National Archives that this feature might be misleading to users. For this reason, the project placed great emphasis on user-experience (UX) design. The presentation and positioning of the links underwent iterative design, testing and refinement to create the version that is now available[3]. The accompanying text is carefully worded to make it evident that the links are not ‘recommendations’ and the feature is clearly badged as ‘BETA’ to indicate that we will continue to refine it.



Other possible matches BETA

The following records may contain information about the person described above. As the links are found by computer analysis, we cannot guarantee they are the same individual or that every record in which the person appears will be listed.

Figure 3. Website text for the ‘Other possible matches’ feature in Discovery

Our process generates numerous links, from high confidence connections right down to tenuous, low confidence associations. Only links above a pre-determined threshold are published in Discovery. This is not entirely satisfactory. As a digital archive creating a digital infrastructure for research, we do not believe that it should be our role to pre-empt the researcher’s level of interest in the data we hold. Our preferred model would be to give the researcher control to tailor thresholds to their own purpose – many will want only high confidence matches while others may wish to explore more possibilities.

For the links themselves, user-testing revealed that expressing confidence as a percentage was not meaningful. Our confidence scores have now been translated to ‘strong match’, ‘possible match’ and ‘weak match’, where a weak match represents a confidence of around 30% (depending on the records in question). This testing drew heavily on our UX specialists, and for future developments we would hope to have this expertise more closely integrated with the project team.

The project has underlined that we cannot treat user experience design or evaluation as an afterthought, even in primarily analytical projects such as TTT. Computational approaches are inherently uncertain and our investigation of these techniques is starting to highlight the uncertainty already present in our existing descriptions. As we learn to embrace this uncertainty, usable interfaces and effective interpretation will be key to ensuring that new features are well-tailored to users’ needs and that the potential benefits can be realised by the research community.

F. *Linking as an enabler for access*

The links created by TTT have three main areas of application for archives and researchers:

Firstly, they support research at an individual scale, by aiding discovery and navigation within and between collections. The features already delivered make this possible to some extent, our main limitation is simply the volume of data that we can prepare and process. Researchers have expressed some demand for access to lower-confidence links than those which have been published via Discovery and we will need new structures and interfaces for publishing data to make this possible.

Secondly, they improve our decision-making on closure of name-rich records, via rigorous linking of personal records with death-registration data. This technique has already been used to accelerate opening for records from two series[4], releasing information about deceased individuals that would otherwise have remained closed for 100 years from their dates of birth [8]. For this purpose, the availability of linking algorithms, robust confidence measures and authoritative reference data are of great importance and we face real challenges in interpreting confidence and communicating risk for decisions based on this type of analysis.

Finally, they support analysis of extensive networks of links to reveal larger-scale patterns within the data. We have so far generated several million links at various levels of confidence. We could publish these as a stand-alone resource, perhaps in the form of GraphML data, but we have recently made great efforts to provide unified access via Discovery, reducing the number of separate services our readers must use. Although a future Discovery may look very different, the principle of unified access is an important one. We cannot yet integrate the large-scale data generated by TTT with our catalogue in such a way that it can be explored and analysed. Enabling this integration will require a very different model for the information we hold about digital records.

III. CHALLENGES FOR ARCHIVES AND IMPLICATIONS FOR ARCHIVAL PRACTICE

The project addressed various technical challenges in its three key areas of investigation: operating over diverse collections, establishing confidence and handling fuzzy data. However, we also encountered challenges relating to our people and the nature of our archival collections.

A. *People*

Computational archival science projects, such as Traces through Time, typically require a wide range of skills and experience that are unlikely to be found within a single heritage institution. The project required expertise in data modelling, statistical modelling, data mining and natural language processing as well as archival science, software engineering and user-experience design plus, of course, an understanding of the needs of researchers. Our approach was to build a research consortium drawing together relevant expertise from academia and the archives sector. Over a two-year period, more than twenty individuals from six institutions have contributed their particular skills, experience and insight to the project. As we start to link to records held outside The National Archives, we expect this network of partnerships to grow, and we hope to be able to share what we have learned more widely across the archives sector.

[3] See discovery.nationalarchives.gov.uk/details/r/C9751641 for an example of the TTT *Other possible matches* feature

[4] Service records of the Durham Home Guard in TNA series WO 409 and the NHS Central Register, (‘1939 Register’) in TNA series RG 101.

Over the course of this work, The National Archives has started to build its own capability and capacity in these disciplines. The project has included broad internal consultation with archivists and records specialists, and has supplied a concrete illustration of some of the challenges we face. This has helped to engage colleagues in discussion about the potential of computational approaches to transform use of our collections. As we increase our reliance on probabilistic approaches to support decision-making and access, we will need to develop ways of explaining these methods for a non-specialist audience, foster statistical thinking and greater 'data literacy' amongst our staff and find ways to understand and express uncertainty for our users.

B. Data

The project has highlighted the need for The National Archives to rethink our drivers for and approach to digitisation. In early digitisation projects, our emphasis (and that of most other archival institutions) was facilitation of remote access, creating digitized assets with the assumption that these will be consumed in much the same way that an on-site researcher reads a paper file. In effect, our traditional digitisation activities deliver a 'picture' of the page, alongside some limited indexing. We have found that digitised resources created in this way do not readily lend themselves to computational analysis and we are beginning to realise that our established approach to digitisation does little to facilitate innovative use of the records. In particular, TTT expended significant effort in re-structuring descriptions to create machine-processable data.

We need to re-define what it means to digitise a record, striking a new balance between our focus on delivering an image to a remote reader and the opportunity to create a dataset to additionally support larger-scale analysis and programmatic investigation.

Our approach to indexing is also important. Manual transcription is expensive and is usually intended only to provide a finding aid, enabling an image to be identified and located. Key fields such as name, and sometimes date of birth, address or service number might be captured. There is little consistency between projects since capture optimises cost/benefit for each individual series of records, with little regard to interoperability with existing datasets or to the future value of the data as a research resource in its own right. We need, at the very least, to capture this information more consistently.

Data elements with seemingly little research value can assume greater prominence when the archive is treated as a big, interlinked dataset. For example, service records often state the name and relationship of the subject's next-of-kin. This information is a low priority for transcription since it is unlikely to provide an access route for search. But, for women's records, this field may give the only indication of whether the recorded surname is a maiden name or a married name – vital information if we are to link successfully to other occurrences of that individual in the records. There will be other examples of attributes whose research value is not obvious until we consider them in the context of other linkable data.

The availability of this type of data has been an ongoing constraint for Traces through Time. We have had some limited success in the application of Optical Character Recognition (OCR) technology to fill this gap. OCR is not in widespread use at The National Archives because the output is felt to be of insufficient quality for inclusion in the catalogue. But, if we recognise 'computation' as an entirely different use-case to 'reading', the techniques can offer real benefit. TTT's probabilistic approach is designed to work with fuzzy data, accommodating variations in data quality. We may find that data that falls below 'catalogue standard' can still be of utility for linking.

Our early results suggest that this application of computational approaches to produce machine-readable text from image data has potential. OCR and the related HTR (Handwritten Text Recognition) require more detailed investigation for this purpose.

IV. IMPLICATIONS FOR ARCHIVES

A. Archival applications of probabilistic techniques

Improved access to computational techniques could transform access to archive collections. Two examples have been trialled for TTT and are described here: use of linking for navigation and to inform closure decisions; and use of automated transcription (OCR and HTR) to make more data available for analysis. Other potential applications include automatically generating descriptions for born-digital records; aiding creating bodies in appraising records and identifying the presence of sensitive content; and applying linking techniques to improve contextualisation for unstructured born-digital content. In each case, there are two challenges that we must overcome: the need for improved computational techniques and the implications for archives of the inherently probabilistic and uncertain nature of these techniques. Addressing the former will require greater collaboration between archivists and computer scientists and joint working on this problem domain. The latter requires a significant shift in archival thinking. We will need to embrace the potential of probabilistic approaches, and develop ways of working within this new paradigm, including new approaches to communication and transparency. For The National Archives, our readers' response to TTT's 'Other possible matches' feature will help inform our thinking on how to communicate and present further automated enhancements to the catalogue.

B. Fluid catalogues

The links generated by TTT provide access to records occurring in very different contexts, possibly held by different institutions and linked only because they concern the same individuals. The same techniques could clearly be applied to draw records together in different ways, through linking other entities.

This disrupts, but need not replace, our traditional emphasis on arrangement by originating body (*fonds* based description) [9]. Digital records often have little inherent order but can readily be arranged according to any number of features present in the records: not only 'metadata,' but thematically, by format or size or according to derived attributes such as links between records. They can be sorted, clustered, copied and moved without disruption to the archived copy (and the concept of the 'original digital record' is, itself, questionable) and there is no reason that we as archivists should elevate any one classification facet (such as the hierarchical structure of the originating body) over other facets for the purpose of providing access.

Releasing this potential may mean more gradated access, recognising 'reading' and 'computation' as distinct, but equally valid, ways of using the record. We could allow the expert to lead the uninitiated whilst offering raw data for the technically literate researcher.

V. The future digital archival catalogue

It seems clear that a future Discovery service will need to provide access to more varied, less rigidly structured information about our archival collections. For born-digital and some digitised records the boundary between record content and metadata is already blurred, and it may become meaningless to maintain this distinction. We will need to offer not only descriptions supplied by creating bodies or curated by archivists, but automatically enhanced descriptions and user-contributed content. Links will provide support for inference. Information will overlap or even conflict and we may see a multiplicity of views about any particular archival resource.

Our understanding of the records may change over time through linking, contextualisation, use and re-purposing of records and as new information emerges. We will need to build this temporal awareness into the digital archive.

Some of this diversity of content already exists, and this potentially chaotic situation is currently addressed through providing access to only a subset of our knowledge about our records, in the form of our archival catalogue. We recognise that much useful data falls outside this narrow definition, but we lack a mechanism for indicating the reliability of other sources or providing meaningful and coherent access.

Although we are beginning to form ideas about the types of use that the future digital archive will enable, and the diversity of information they will manage, we are still some way from understanding how to achieve this or how it might appear to the user. A catalogue data model and technical architecture that would enable all of this would look very different to Discovery today, and we will need new technical approaches to support this, with graph and probabilistic database technology likely to feature in our future solution. Emergent archival standards [10] identify some of these challenges, but may not be disruptive enough to address the full range of issues discussed here and further innovation will be required.

Our way forward lies in acknowledging that our understanding of the records is imperfect and in enabling access to this information, in all its diversity. Our descriptive practice must evolve, giving more control to the user and opening up the archive to new, natively digital, research methods. This is a fundamental shift from our first-generation lists and images, to a second-generation probabilistic and temporally aware platform for managing and publishing archival data.

Acknowledgment

I would like to thank the Arts and Humanities Research Council and The National Archives for their support of the Traces through Time project; Professor Jane Winters (University of London), Dr Roger Evans (University of Brighton) and Dr Arno Knobbe (University of Leiden) who were my Co-Investigators in the initial phase of this project; and my colleagues Mark Bell, Alex Green and John Sheridan for their constructive review of an earlier draft of this text.

References

[1] V. Johnson, S. Ranade, D. Thomas, "Size matters: The implications of volume for the digital archive of tomorrow – a case study from the UK national archives", Records Management Journal, Vol. 24 Iss: 3, 2014, pp.224 – 237, doi: 10.1108/RMJ-01-2014-0004.

[2] International Council on Archives, Subcommittee on descriptive standards, "ISAD(G): General International Standard Archival Description", Second edition, Sept. 2011, www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

[3] The National Archives, "Digital strategy", unpublished.

[4] The National Archives, "Archives inspire: The National Archives plans and priorities 2015−19", 2015, www.nationalarchives.gov.uk/ documents/archives-inspire-2015-19.pdf.

[5] International Council on Archives, Subcommittee on descriptive standards, "ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families", Second edition, Sept. 2011, www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf.

[6] schema.org, "Thing > Person", version 3.1, schema.org/Person.

[7] The National Archives, "Traces through Time person schema", github.com/nationalarchives/traces-through-time.

[8] S. Ranade, "Durham home guard pilot: linking with death register data", The National Archives internal report, 2012, Unpublished.

[9] H. Jenkinson, "A manual of archive administration", Second edition, London : P. Lund, Humphries & co., ltd., 1937.

[10] International Council on Archvies, Experts group on archival description, "Records in contexts: a conceptual model for archival description", Consultation Draft, v0.1, Sept. 2016, http://www.ica.org/sites/default/files/RiC-CM-0.1.pdf