WITN09960100

Witness Name: Paul John Tombleson
Statement No: WITN09960100
Dated: 22 August 2023

POST OFFICE HORIZON IT INQUIRY

FIRST WITNESS STATEMENT OF PAUL TOMBLESON

I, PAUL JOHN TOMBLESON, of 15 Canada Square, London E14 5GL say as follows:

Introduction

1. I am a Partner in the Forensic practice at KPMG LLP ("**KPMG**") specialising in Forensic Technology. I have led many eDisclosure matters for clients in the UK and globally since 2005.

2. The facts in this witness statement are true, complete and accurate to the best of my knowledge and belief. Where I refer to my beliefs, these beliefs, and my knowledge contained within this statement, are informed by my colleagues at KPMG, including in particular KPMG's Forensic Technology team. I have been assisted in preparing this witness statement by KPMG's Office of General Counsel, advised by Linklaters LLP.

3. This witness statement has been prepared in response to a request by the Post Office Horizon IT Inquiry (the "**Inquiry**") pursuant to the Inquiry Rules 2006, dated 31 July 2023 (the "**Rule 9 Request**"), relating to KPMG's engagement to provide eDisclosure services to the Post Office Limited ("**POL**") and the three disclosure issues which have been identified during the Inquiry. We understand

WITN09960100

the Rule 9 Request to be referring to the following topics: (i) search terms; (ii) family documents; and (iii) deduplication.

4. The role of KPMG in the Inquiry is as an eDisclosure provider, supporting POL to respond to requests related to the Inquiry. This includes taking instruction from POL's relevant external solicitors, primarily Herbert Smith Freehills ("**HSF**"). We were engaged by POL in January 2021 and I am the Partner leading the KPMG team, which comprises approximately 26 people as at the end of July 2023.

5. There are five main components to the eDisclosure services provided to POL by KPMG:

   - Data scoping, collection and migration;
   - Data preparation;
   - Data processing;
   - Review and hosting using Forensic Technology Tools; and
   - Disclosure.

6. As at the end of July 2023, KPMG hosts over 29 Terabytes of data for POL across 19 workspaces in a database called Relativity (the "**Relativity eDisclosure Database**"). There are four main workspaces related to the Horizon IT Inquiry containing approximately 60 million individual documents.

WITN09960100

The events which led to the three disclosure issues

7. I became aware of the late disclosure to the Inquiry of a document from reports in the media. This document was entitled 'Appendix 6 – Identification Codes' ("**Appendix 6**") (an example of Appendix 6 is Inquiry URN **POL00115674**). Mr Foat, POL's General Counsel, gave evidence on the three disclosure issues on 4 July 2023. KPMG was not involved in providing information to support Mr Foat's witness statement dated 21 June 2023 or his oral evidence.

8. At POL's request, we conducted an internal KPMG review (the "**Internal Review**") into the events that led to the disclosure issues and this included a review of the circumstances around the delayed discovery and disclosure of Appendix 6 to the Inquiry.

9. In the following sections, I address each of the three issues.

Search terms

10. It is commonplace for search terms and keywords to be used in eDisclosure to narrow down the population of documents requiring review. Documents that contain a search term are referred to as responsive documents or 'hits'.

11. Typically, on receipt of a new request from the Inquiry, HSF notifies KPMG and provides search instructions (which include the data sources, custodians, date ranges, search terms and other search criteria). KPMG has not had any involvement in the development of the search criteria or search terms arising from a request, aside from providing ad hoc technical assistance with the

WITN09960100

searching and Boolean logic to be deployed in Relativity. Boolean logic is a way of structuring searches using different operators (e.g., AND, OR) to get the most precise results.

12. In certain instances, Appendix 6 is in a family group of documents with the document 'Appendix 3 – Guide to the Preparation and Layout of Investigation Red Label Case Files. Offender reports & Discipline reports' ("**Appendix 3**") (an example of Appendix 3 is Inquiry URN **POL00038452**). I understand from Mr Foat's witness statement of 21 June 2023 that Appendix 3 and Appendix 6 are documents that appear to be related to the Inquiry's requests 11 ("**R9(11)**") and 14 ("**R9(14)**").

13. Based on the findings from our Internal Review, I can confirm that Appendix 6 was not responsive to the search terms prepared by HSF for R9(11), dated 28 February 2022.

14. However, Appendix 3 was responsive to three identical R9(11) searches which HSF instructed KPMG to deploy in March / April 2022, resulting in 19 versions of Appendix 3 being reviewed and tagged by HSF between 28 March and 16 May 2022.

15. I note that neither Appendix 3 nor Appendix 6 appears to have been responsive to the initial search terms developed by HSF in 2022 for R9(14), dated 15 June 2022.

16. To the best of my recollection, KPMG has not had any conversations with POL about search terms to be applied on specific requests. KPMG's instructions came from POL's external solicitors only.

WITN09960100

Family documents

17. Many documents that are hosted in the Relativity eDisclosure Database are part of a family. A parent-child relationship exists for those document families. For instance, an email (the parent) with an attachment (the child).

18. Typically, search terms will identify one or more individual documents that are part of a family, but not every family member. Using the example above, the attachment to the email may be responsive to the search terms whilst the parent email is not.

19. In the vast majority of cases, the instructions that KPMG has received (prior to recent changes following the identification of the disclosure issues) were to provide only the responsive documents for review, i.e., only those members of a family with a direct search hit. This means that the reviewer only had to make a decision on a specific document responsive to the search terms, albeit a reviewer always had the ability to examine family documents in Relativity.

20. I am not aware of POL's instructions to its external solicitors regarding the review of families of documents.

21. All 19 versions of Appendix 3 referred to in paragraph 14 were responsive to search terms. These responsive documents were reviewed by HSF and given a relevance decision.

WITN09960100

22. Eight of the 19 versions of Appendix 3 were in the same family group as Appendix 6. None of the family documents for these eight were reviewed (until May 2023).

23. To the best of my recollection, KPMG has not had any conversations with POL about the review of family documents. KPMG's instructions came from POL's external solicitors only.

Deduplication

24. KPMG is hosting a wide variety of data from many different data sources in the Relativity eDisclosure Database. This includes email data from various individuals' mailboxes, SharePoint files, CDs, backup tapes and hard copy scanned documentation. Within this data set are a significant number of duplicates.

25. From a Forensic Technology perspective, duplicate documents are those with the same MD5 Hash algorithm ("**MD5#**") number. The purpose of an MD5# number for documents is to provide a unique and fixed-size fingerprint, in the form of a 32-digit hexadecimal number string, that represents the content of the document. It is commonly used for detecting when changes have occurred to a document.

26. Deduplication based on MD5# numbers is an important eDisclosure technique to reduce the number of documents prior to review. It is possible to deduplicate at an individual document level ("**item level deduplication**") or at a document family level ("**family level deduplication**").

WITN09960100

27. I have provided an example below of each form of deduplication based on five document families (Family 1-5). Four of the document families have a different parent email (Family 4 and 5 have the same parent email). All five families contain the same Attachment 1, four families contain the same Attachment 2, three families contain the same Attachment 3, and two families contain the same Attachment 4.

28. In this example, search terms are run across the data set and only Attachment 1 (highlighted in red) is responsive to those terms. Therefore, there are five responsive documents across the data set, since Attachment 1 is a child in every family.

Table 1: Deduplication example

Family 1	**Family 2**	**Family 3**	**Family 4**	**Family 5**
Email 1	Email 2	Email 3	Email 4	Email 4
Attachment 1	Attachment 1	Attachment 1	Attachment 1	Attachment 1
	Attachment 2	Attachment 2	Attachment 2	Attachment 2
		Attachment 3	Attachment 3	Attachment 3
			Attachment 4	Attachment 4

29. KPMG is routinely asked by POL's external solicitors to report statistics on the number of hits and (in some cases) the number of hits and families. In the example above, there are five hits and 19 hits with full families.

WITN09960100

30. If family level deduplication was applied across the dataset, the MD5# of the document families is compared and duplicates are removed. Since two families are identical (Family 4 and 5), after family level deduplication there would be four families and 14 documents remaining in the data set for review.

31. If item level deduplication was applied across the hits, the MD5# of the responsive items (i.e., Attachment 1) is compared and duplicates are removed. Since all five versions of Attachment 1 are identical, after item level deduplication there would be only one version of Attachment 1 (and its family) promoted for review.

32. In this example of item level deduplication, it is important to understand which version of Attachment 1 is deemed to be the 'master' and which four versions are deemed to be 'duplicates', since four of the five families are different. If the version of Attachment 1 in Family 1 is deemed to be the master, this document (and its family) will be promoted for review. However, the reviewer would not see Family 2, 3, 4 and 5 since they contain duplicate versions of Attachment 1, even though their families are different.

33. KPMG uses a Relativity 'Update Duplicate Status' script as the basis for item level deduplication using MD5# values. The script tags documents as either master, unique or duplicate in order to filter out duplicates and pass only master and unique documents, and their families, for review. There are a variety of different ways to sort duplicates but the default setting is to use Artifact ID (Artifact ID is a metadata field populated by Relativity when a document is loaded into a new workspace). Hence, the first document loaded in the workspace becomes the master.

WITN09960100

34. A summary of the number of documents requiring review based on the above example is as follows:

Table 2: Review scenarios – number of documents requiring review

	Family level deduplication	**Item level deduplication**
Review of full families	14	2* *Assuming Family 1 is deemed the 'master'
Review of responsive documents	4	1

35. There are risks and implications from both types of deduplication technique, with the main difference being the trade-off between the number of documents requiring review versus having the context from the review of full families.

36. As noted in paragraph 19 above, HSF's review strategy appears to have been primarily focussed on responsive documents and not their full families. We understand that this was in an effort to identify a review population that was reasonable to review within the timescales required by the Inquiry.

37. In line with this approach, where deemed necessary, HSF has instructed us to deduplicate between responsive documents using MD5#, i.e., item level deduplication. We therefore undertook deduplication at a responsive document

WITN09960100

level, rather than a family level, to reduce the number of documents requiring review.

38. Deduplication has not been applied as standard, but only on certain requests or sub-questions of requests, when instructed to do so by POL's external solicitors. MD5# item level deduplication first took place in relation to request 10 ("**R9(10)**") to assist HSF to reduce the number of hits requiring review.

39. Based on the findings from our Internal Review, in relation to the three identical R9(11) searches that KPMG ran and the 19 versions of Appendix 3 reviewed in March, April and May 2022:

    - The first search was run on 25 March 2022. 16 versions of Appendix 3 were responsive to the search terms. Item level deduplication was instructed to be applied to these documents and nine duplicate versions of Appendix 3 were removed before review. The seven remaining versions of Appendix 3 each had a different MD5# and were selected as having master status based on the earliest Artifact ID (as described in paragraph 33 above). These seven versions were provided to HSF for review and tagged as 'not relevant' on 28 March 2022. These seven versions comprised five standalone documents and two families. The two families were partially reviewed by HSF, but neither of the two families contained Appendix 6.

    - The second search was run on 5 April 2022. The same 16 versions of Appendix 3 were responsive to the search terms as the first search. No item level deduplication was instructed for these documents and,

WITN09960100

therefore, the nine unreviewed documents (that were removed as duplicates in the first search) were provided to HSF for review:

- Eight versions of Appendix 3 were tagged as 'not relevant' by HSF on 8 April 2022. HSF did not review the families of these responsive documents at that time. These eight versions were in the same family as Appendix 6 and thus Appendix 6 was available for review in Relativity as described in paragraph 19 above; and
- The ninth version of Appendix 3 was tagged as 'relevant' by HSF on 9 April 2022. HSF did not review the family of this responsive document at that time. This version was not in the same family as Appendix 6.

- The third search was run on 11 May 2022. This search was run following rectification of an issue that had arisen during the migration of a database of documents to KPMG from a previous eDisclosure service provider to POL. A further 21 versions of Appendix 3 were responsive to this rerun search, 14 of which were in the same family group of documents as Appendix 6. In line with the March 2022 search instructions from HSF, deduplication was applied (based on the earliest Artifact ID):
  - 18 versions of Appendix 3 were not made available for review by HSF as they were MD5# item-level duplicates of versions already reviewed by HSF in March and April 2022; and

WITN09960100

- Three versions of Appendix 3 were provided to HSF for review and tagged as 'not relevant' on 16 May 2022. All three versions were standalone documents without associated families.

40. To the best of my recollection, KPMG has not had any conversations with POL about deduplication techniques. KPMG's instructions came from POL's external solicitors only.

What steps have been taken to remediate the issues

41. KPMG is providing significant support to POL and its external solicitors with the ongoing remediation exercise.

42. We have identified the specific searches and related Inquiry requests where item level deduplication took place. We have identified 11 affected requests.

43. We are helping, under the instruction of HSF, to identify, provide for review and produce additional documents to the Inquiry related to:

    - The 'Policy Review' Searches Remediation and Assurance workstreams: Considering any 'gaps' in search terms previously used and applying additional keywords across certain data sets;

    - The Families Remediation workstream: The review of all unreviewed and / or unproduced families of documents already produced; and

    - The Deduplication Remediation workstream: The review of families of duplicates of documents already produced.

WITN09960100

Systems and processes in place to avoid future issues of a similar nature

44. The specific circumstances that led to the late disclosure of Appendix 6, based on our Internal Review, are set out above and primarily relate to search terms and review of family items.

45. We are continuing to support POL's external solicitors to apply search terms and search criteria across the Relativity eDisclosure Database, but we are not involved in their development.

46. We are providing POL's external solicitors with responsive items and their full families for review, and we will continue to do so unless instructed otherwise.

47. I also recognise that there are risks and implications from the selection of a deduplication technique, as the example in paragraphs 27 to 35 above shows. Since 21 June 2023, HSF has instructed KPMG to only use family level deduplication, and not item level deduplication. In line with these instructions, we have not performed item level deduplication since this date. Any deduplication going forwards will continue to be in line with the instructions of POL's external solicitors.

WITN09960100

Statement of Truth

I believe the content of this statement to be true.

Paul John Tombleson

Signed: GRO

Dated: 22 August 2023

WITN09960100

Index to the First Witness Statement of

Paul John Tombleson

No.	URN	Document Description	Production Number
1	POL00115674	Appendix 6 – Identification Codes	POL-0115834
2	POL00038452	Appendix 3 – Guide to the Preparation and Layout of Investigation Red Label Case Files – Offender reports & Discipline reports	POL-0027763