A Tale of Two Traces – Archives, Diplomatics, and Digital Forensics

Fred Cohen – CEO – Management Analytics Senior member, IEEE, Pebble Beach, CA 93953

Abstract—**This paper focuses on two examples of legal matters involving archived data, one a digital archive of born-analog data, and the other a digital archive of born-digital data. Their resolution is explained, and along the way, several of the challenges and issues related to digital archives, the transition from classical diplomatics to modern diplomatics, digital forensics in the light of current record-keeping systems, and related facts and supporting data points are explored.**

Keywords-component; questioned digital documents; digital diplomatics, archival science, digital records management.

I. Introduction and Background

In October of 2013, a house was sold, and in the closing days, there was a dispute about square footage that could have stopped the sale or substantially changed the terms of the agreement. The issue was only ultimately resolved by examining documents from city and county digital archives of born-analog data. We will call this Case 1.

In an unrelated legal dispute, two ex-partners were in stark disagreement regarding whether one of the partners undertook a competitive business while the partnership was still operating, and thus acted in bad faith regarding the partnership. The issue was brought to light because of the content of a Web-page depicted on archive.org's Wayback Machine, asserted to be an archive of Internet Web sites from historical (since the mid-1990s) time frames. We will call this Case 2.

The issues in these cases are not new. In fact, they are very old, and by looking back to the history of archives, diplomatics, and related aspect of records management, insight may be gained about futures in these fields.[1]

A. Causality as a foundation of science

Foundational to science is the notion of causality. Cause (C) acts through (→) mechanisms (m) to produce effects (E). Expressed as $C \rightarrow^{m} E$, this forms the basic assumption of science as a whole and scientific evidence in the narrow sense of legally admissible evidence. Noteworthy are the notions that correlation is not causality and that effect does not imply cause.[2] In order to form a scientific hypothesis about a legal matter, a hypothesis of a mechanism by which cause produced effect must be formed, with the effect being the traces found and the cause being a hypothesized act of interest to the matter. History suggests progress in the regard.

B. Diplomatics

The field of diplomatics is often identified as founded in 1681 when the famous French phililogist[3] Mabillon[4] first published the results of an analysis of approximately 200 documents, divided into categories and examined with regard to material, ink, language, script, punctuation, abbreviations, formulas, subscripts, seals, special signs, chancery notes, and so forth. He created descriptions to allow the detection of forgeries and identified ground truth based on recurrence of intrinsic and extrinsic elements in documents from same time and place.[5] In modern terminology, and taking some liberties in usage, he used redundancy to test for consistency. Note that this approach was based on correlation, but causality was also present in the form of known chanceries or scriptoria traditions (cause) and capabilities of scribes over the ages (mechanism). In addition, and perhaps more vitally, no ground truth was available for much of this effort because the documents were too old for eye witnesses and the documentary evidence supporting the claims was in question along with the claims.

Diplomatics and the approaches it used formed much of the basis for admissibility of evidence and the establishment of criteria for evaluating evidence still used today. This field, combined with its principles in application, is still used today for questioned document analysis, and by extension, its principles and many of its methods are in use or have analogous use in digital forensics. The modern and historic reconstruction of causes acting through mechanisms to produce effects forms an experimental basis for diplomatics, except, of course, that accelerated aging and similar methods are approximations or models.

C. Archival science, archives, and public records

The field of archival science emerged over time as part and parcel of the need to keep reliable public records, for example, of land ownership. Ancient record-keeping systems date back as far as history, and indeed much of history is based on the records ingested into and retained by archivists in the archives of different administrative bodies. In the legal system, public records are generally admissible for the truth of what they self-indicate, and are presumed trustworthy (i.e., reliable[6], authentic[7], and accurate[8]) in the legal system,

1 This paper has definitions, statements, and details that are part of cited works that, if cited in each instance, would use more space in citation than content. The interested reader should review cited documents for the detailed basis.

2 F. Cohen, Digital Forensic Evidence Examination, ASP Press, 2008-14 has a more detailed exposition on this.

3 Expert in the analysis and transcription of documents

4 Dom Jean Mabillon, "De Re Diplomatica", 1681, Saint-Maur, France. As referenced in Duranti

5 See L. Duranti, "Diplomatics – New uses for an old science", Scarecrow Press, 1998, ISBN 08-108-352-82 for many more and more definitive details.

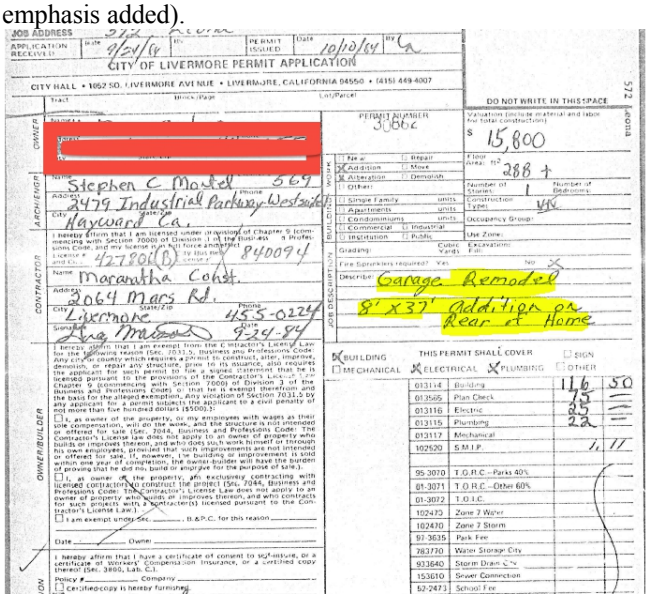
6 Reliable: the record is a true statement of fact

7 Authentic: the record has not been corrupted or tampered

8 Accurate: truthful, exact, precise, or complete

when properly introduced and marked with appropriate seals, signatures, and/or special signs from the legal entities that produce them. While diplomatic analysis may be used to try to refute these records and/or to rehabilitate them after attempted refutation, they are generally trusted, and built and maintained in such a manner as to reasonably justify this trust. At least this is their nature in the analog records space.

This analytical approach is based on a set of redundant acts by independent trusted actors forming a set of archival fonds[9] associated with different archival units, programs, or institutions. These records and fonds include explicit formal elements designed to provide assurance that records are authentic, accurate, and reliable over their life. This is undertaken by providing a chain of custody in a transparent system of record-keeping through redundant information associating acts related to the record with the record in the context of the fonds, and the fonds in context of the archives. This is sometimes called the archival bond.[10] Paper records are often annotated over time, marked with stamps, altered with updates and changes, and so forth. As an example, Figure 1 shows a (redacted) document provided in response to a request for a legal record related to ownership (yellow emphasis added).


CITY OF LIVERMORE PERMIT APPLICATION

CITY HALL • 1052 SO. LIVERMORE AVENUE • LIVERMORE, CALIFORNIA 94550 • (415) 449-4007

PERMIT NUMBER 30862

$ 15,800

Stephen C Mastel 569

2479 Industrial Parkway West

Hayward Ca.

840094

Maranatha Const.

2064 Mars Rd.

Livermore 455-0224

9-24-84

Garage Remodel

8' X 37' Addition at Rear of Home

THIS PERMIT SHALL COVER: BUILDING, ELECTRICAL, PLUMBING, MECHANICAL

013114	Building	11.6 50
013505	Plan Check	7.5
013116	Electric	2.5
013115	Plumbing	2.2
013117	Mechanical	
102920	S.M.I.P.	1.11
95-3070	T.O.R.C.--Parks 40%	
01-3071	T.O.R.C.--Other 60%	
01-3072	T.O.I.C.	
102470	Zone 7 Water	
102470	Zone 7 Storm	
97-3635	Park Fee	
783770	Water Storage City	
933640	Storm Drain City	
153010	Sewer Connection	
62-2473	School Fee	

Figure 1 – An example from Case 1

As can be seen from Figure 1, a series of writings appear on the form, and over time, additional writings are added to the record. The record, in this case, is a dynamic document that is updated to reflect officially authorized changes, as shown by the seals of the officials carrying out those acts. As the record evolves over time, it is retained in a chain of custody and supported by the fonds in which is resides, which reflects dates of access and related information. As a legal document, this is considered proof of the facts contained therein and is inherently regarded as reliable and accurate based on the source, and authentic based on being supplied to the court with proper form and seals intact.

This document might have been moved into an archival repository for a period of time and returned to active use later, and such movements will be annotated on the document and/or within the fonds in which it resides over time. The redundant information from these various sources (e.g., transmission from use to archives and back as signed by parties on each side, placement in the fonds in sequence over time, markings on cover sheets and/or envelopes and/or the documents themselves, etc.) can be examined by the diplomatics expert in the context of the methods used by the record-keeping system to make a determination of authenticity or to refute or challenge the presumption of authenticity based on a lack of adequate evidence of, or evidence of inadequate, custody and control.

D. Digital records

In the digital records space, many of the methods that made analog records reliable over time were not translated into the new forms of record-keeping. For example, the processes involving purchases of property often included the use of an actuarial working for or on behalf of the parties and authorized by government to certify that an individual identified by a government identification signed a document in the presence of the notary. But increasingly digital systems allow digital signatures not even using the individuals own hand. Rather, the self-identified individual agrees electronically over the Internet to adopt a signature form for use in signing documents. The documents are sometimes incorrectly presented (i.e., with incorrect data fields), with the results produced as digital documents reflecting different (and in some cases corrected) content than what was actually presented for signature.[11] This is how deals are now done.

Such record-making and -keeping systems are potentially enormously problematic in legal terms, but are not often challenged, or have not yet been so challenged. They do not guarantee that what is agreed to is what is presented, they include and present false information and change it after agreement, don't provide a copy in the form of original[12], an imitative copy[13], or even a simple copy[14] to the signatories, and don't use actual signatures traceable to the individual or demonstrably different from other "adopted" signatures. Documents may be presented differently than they ever appeared before (i.e., as a pseudo-original[15]), even when and if they are ultimately presented and/or submitted in court as authentic, reliable, and accurate.

---

9 Fonds: the aggregation of documents that originate from the same source

10 Archival bond: the relation between a document and the previous and subsequent ones produced in the course of carrying out a business matter

11 This was a fact pattern in Case 1, but was not related to a disputed issue in this particular case.

12 A 'copy in the form of original' is identical to the original in all respects, but is issued after the original.

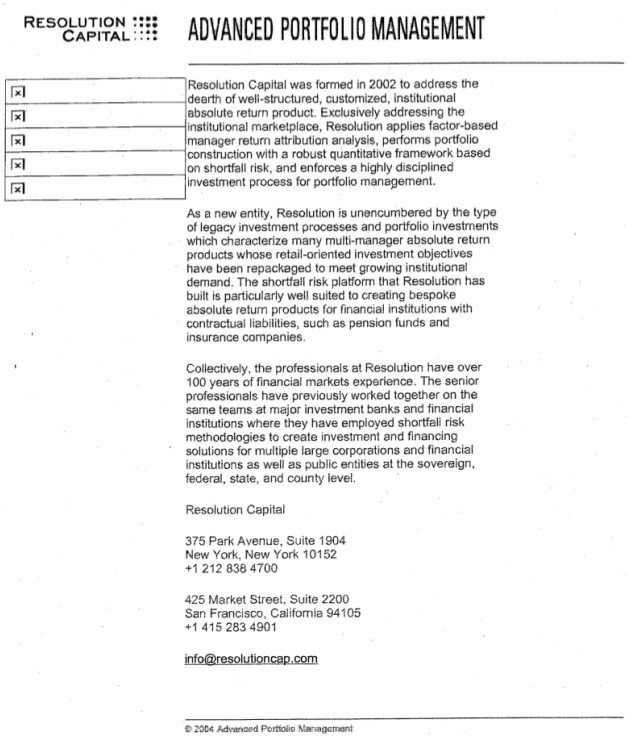
13 An 'imitative copy' is a reproduction of both the form and content of a record.

14 A 'simple copy' only transcribes the record content.

15 A 'pseudo-original' has the pretense of originality

These pseudo-original documents are then declared as public records, and from that point forward, recognized, treated, and presumed as authentic renditions of contracts. They become part of corrupt and inauthentic digital records and eventually make their way into the archived and permanent records of societies. The metadata associated with these records often lacks fields required by record management systems and archives, if present they may be incorrect, the mechanisms are not transparent, and they not available to the individuals forming the contract.

Figure 2 shows an example of a presentation made as part of the collection of potential evidence in Case 2. This depiction of a digital record reflects what, in some jurisdictions, is legally admissible as an archival document and may be given the presumption of authenticity, reliability, and accuracy. In particular, note that "Resolution Capital" appears with "Advanced Portfolio Management" together on the page. This is a depiction saved from a screen image of what was seen at the time one of the parties gathered what they believed to be evidence in support of their case.


RESOLUTION CAPITAL

ADVANCED PORTFOLIO MANAGEMENT

Resolution Capital was formed in 2002 to address the dearth of well-structured, customized, institutional absolute return product. Exclusively addressing the institutional marketplace, Resolution applies factor-based manager return attribution analysis, performs portfolio construction with a robust quantitative framework based on shortfall risk, and enforces a highly disciplined investment process for portfolio management.

As a new entity, Resolution is unencumbered by the type of legacy investment processes and portfolio investments which characterize many multi-manager absolute return products whose retail-oriented investment objectives have been repackaged to meet growing institutional demand. The shortfall risk platform that Resolution has built is particularly well suited to creating bespoke absolute return products for financial institutions with contractual liabilities, such as pension funds and insurance companies.

Collectively, the professionals at Resolution have over 100 years of financial markets experience. The senior professionals have previously worked together on the same teams at major investment banks and financial institutions where they have employed shortfall risk methodologies to create investment and financing solutions for multiple large corporations and financial institutions as well as public entities at the sovereign, federal, state, and county level.

Resolution Capital

375 Park Avenue, Suite 1904
New York, New York 10152
+1 212 838 4700

425 Market Street, Suite 2200
San Francisco, California 94105
+1 415 283 4901

info@resolutioncap.com

© 2004 Advanced Portfolio Management

Figure 2 – An example from Case 2

E. Digital diplomatics

As a field, digital diplomatics today is taking on two meanings.[16] One meaning is the use of digital computing methods to support classical diplomatics. For example, digital methods are being used for word and phrase analysis to detect changes in scribes and to track scribe usage of terms over time so as to date documents more accurately than was previously feasible with manual techniques.

The other meaning is the use of diplomatics methods to authenticate digital documents, and that is the meaning of interest to the present paper:

> Georges Tessier : « On peut donc avancer que la critique diplomatique est née dans le prétoire ou sur le forum à l'occasion de débats judiciaires ou de controverses politiques ou religieuses, quand le nœud du litige ou de la polémique était constitué par un document ou une série de documents contestés ». Cette citation est tirée de L'Histoire et ses méthodes (La Pléiade, 1961) dont Georges Tessier a signé le chapitre « Diplomatique ».[17]

In this context, the relevance of digital diplomatics may be reasonably explored relative to digital forensics, where forensics is from the literal "forensis"[18]

F. Forensic science

Forensic science is often cited as coming to clarity through the work of Locard.[19] Locard identified that when objects come into contact, they each leave parts (traces) of themselves with the other. The mechanism of objects coming into contact leaving traces is called "transfer", and thus we have the scientific notion of causality fulfilled by contact (cause) acting through transfer (mechanism) to produce effect (traces). Traces may be humanly observable (e.g., chunks of rock or mud) or "latent" (observable only through the use of tools as in dust or microscopic particles). Locard undertook studies showing layers (e.g., of mud or dust) indicating sequences of places visited (e.g., layers of mud on shoes) when people transited a city, and associating the transferred traces to locations based on unique properties (e.g., strands of a particular wool from the only factory that produced it in that form in the city).

As forensic science has moved forward, many methods have been developed based on the concept of transfer, as well as methods from the earlier diplomatics area, such as tool mark analysis.

16 This reflects presentations at the Digital Diplomatics conference in Paris, France, 2013-11-14 to 16.

17 http://www.marieannechabin.fr/ - 2013-11-11 blog of Marie Ann Chabin - (English by Google: "It can be argued that the diplomatic criticism is born in court or forum on the occasion of judicial proceedings, or political or religious controversy when the crux of the dispute or controversy consisted of a document or series of documents in dispute." This quote is from The History and Methods (The Pleiades, 1961) where Georges Tessier signed the chapter "Diplomatics".

18 http://www.merriam-webster.com/dictionary/forensic "belonging to, used in, or suitable to courts of judicature or to public discussion and debate" - Latin forensis public, forensic, from forum forum - First Known Use: 1659

19 E. Locard, "The Analysis of Dust Traces", Revue International de Criminalistique I. #s 4-5, 1929, pp 176-249, (translated into English and reprinted in 3 parts in A, J. Police Science, 1930 in V1#3, May-Jun 1930, pp276-298, V1#4 Jul-Aug 1930, pp 401-418, and V1#5 Sep-Oct 1930, pp 496-514.)

G. Digital forensics

As digital systems came into widespread use, the legal system had to deal with evidence in the form of traces of activities within and between those systems. The study of digital traces relative to the legal system was identified as digital forensics, however, as in digital diplomatics, another meaning is used. Digital forensics is also used to describe activities associated with investigation of events in the digital arena, a much broader field closely related to detection and response regimens in computer security.

As a fundamental notion, it has lately been recognized that digital evidence is still trace evidence, is almost always latent in nature, and is **not** transfer evidence. Rather than transfer in the sense of Locard, digital evidence is formed from traces produced by the mechanisms of digital systems operating, typically as stored state from finite state machines that transform state and input into next state and output.

In the term of art from diplomatics, digital traces are produced by transmission[20] rather than transfer. In the digital arena, transmission producing traces is typically also transmission in the sense of electromagnetic, optical, sonic, or other emitting and reception of signals. That is, events in one context produce signals that are sensed in another context and memorialized in the form of optical patterns, configurations of particles, magnetic orientation, or whatever traces the transmission or fixation media supports.

In addition, in the digital arena, the latent nature of evidence is such that a copy in the form of original almost never actually occurs. Rather, a sequence of bits represented in fixed form in/on a medium may be reproduced (at the bit level), while presentation in human readable form is normally an imperfect reflection of the original documentary form. For example, when the information associated with a digital record (e.g., a financial transaction) is originated, the form of entry (e.g., a Web-based entry of a purchase form) is typically very different from the form of transmission (e.g., a series of datagrams sent over the Internet as waveforms in a transmission media), storage (e.g., a sequence of bits stored in a database storage area of a disk drive or in a positive feedback loop in active memory), and presentation (e.g., a line item in a bank statement or an entry in a spreadsheet downloaded by an accountant from the financial institution and used for tax purposes). These notions are not widely recognized or stated in digital forensics today, even though they are certainly always present.

The notions of authenticity, accuracy, and reliability are always at issue in the digital forensics arena relative to the classical notions of documentary form, and the notion that using methods such as cryptographic checksums to verify a lack of alteration of a bit sequence doesn't even begin to address the issues of authenticity of a record in presentation and reliability in the sense of relationship to original writing or any sort of ground truth. Causality works differently.

20 Transmission in the sense of records management and archival science includes physical movement from place to place and the logical handover of control of records without physical movement.

H. Summary of results and issues to follow

In the following sections, distinctions, if they exist, will be identified with classic diplomatics, classic forensics, digital forensics, and digital diplomatics, in the realm of questioned document examination. In parallel with this exposition, the case studies will be examined. These cases are not large or important on their own, but rather reflect the many every day legal issues that naturally occur in human interactions and sit at the heart of how people interact with the legal issues we face in these areas. Finally, we resolve the cases at hand, and draw conclusions. It is our view that questioned digital document examination represents a fusion of diplomatics and forensics. It may reasonably be called digital diplomatics and/or questioned digital documents, without reasonable differentiation. We believe that the reconciliation of these two fields in this arena represents a historic merger and unification of the respective concepts and fields of study.

II. Digital diplomatics vs. forensics vs. Digital forensics

A. Case 1 background

Case 1 involved a dispute over square footage of a house. The seller claimed the same square footage they purchased the house at and as reflected in taxes paid over the duration of their ownership and for some unknown period prior to the earlier purchase. The buyer, a civil engineer, upon assessment, received a different square footage from the inspector's report, proceeded to do their own measurement, and produced yet a third square footage result.

If left unsettled or settled in various ways, this situation could lead to charges of fraud, damage to reputation, a price change of tens of thousands of dollars, retroactive tax readjustment, and delayed- or non-closure of the sale. None of these situations were in the interest of any of the parties, and the settlement of the dispute rested upon documentary evidence in the form of records from various sources, including the prior sale documentation, tax documentation, and city and county records from remodeling, permitting, and inspections. The measurements themselves were also at issue because different measurements (e.g., inside dimensions, outside dimensions, “livable” space, permitted use areas) are based on different definitions in different overlaid jurisdictions (taxation, county building, and city building).

B. Case 2 background

Case 2 involved a dispute between ex-partners in a financial business. The business failed and each went their own way seeking to start a new financial business, with ownership of a domain name remaining with one of the partners. Several years later, in viewing what was believed to be an image of the prior Web site using the Wayback machine at archive.org, the party not retaining control of the Web site was unhappy to find that, according to the displayed content, the subsequent company advertised the new company prior to the termination of the partnership,

leading to the charge of misappropriation of resources, customers, and business from the partnership and failure to faithfully fulfill fiduciary and other duties as a partner.

In this case, the dispute was, at its essence, based upon the form and appearance of the document (i.e., the depicted Web site) as seen in the archival site by the distant user of that site. The depiction was clear as could be. A date selected by the user and indicated in the URL at the top of the Web browser page showed content from the prior business simultaneously displayed in a single Web page with material from the subsequent business. If the depiction reflected reality, there could be little question that a case could be made. The only case that could reasonably be made by the accused party involved questioning the document presented by the 'archive'.

C. A legal view of admitting these documents

While the subtleties of an "Internet archive" vs. other sorts of archives and the question of how to resolve seemingly inconsistent information from different official records may be vitally important to the issues at hand in these cases, there seems no question that, on its face, these documents would normally be admitted in legal proceedings.

The WayBack Machine is a form of automatic storage, while archives 'preserve'. Preservation is a process in which the archivists identify, authenticate, protect, describe, build retrieval systems, provide access to, and otherwise act to protect the material being archived. The term "Internet Archives" in the context of the WayBack machine is a misuse of the term of art 'archive'. Of course people have trusted anything called archives for centuries, and those at archive.org demonstrated excellent marketing skills in using that term.

The legal status of government documents is normally that they are admitted and presumed reliable, authentic, and accurate. Thus the documents supplied in Case 1 operate under this legal presumption.

The 'Internet archive' is a bit more nebulous in that it is a Web site operated by a non-profit (i.e., public interest) corporation, seemingly like a museum or other archive. However, this is what the WayBack machine is **not**. It is not like a museum or an archive because there is no curation or assurance of protection and permanent authenticity from the moment of acquisition.

Ancient documents are normally admitted under the presumption that they were not forged in advance in anticipation of some future litigation that could not have been anticipated by the archivists. The question of how old is old enough to be ancient aside, a strong case can be made that, in this case, the Wayback machine was not operating intentionally to create a forgery, and no claim was asserted that information it stored was altered in any nefarious way. The presumption for such documents is, de-facto, also that of being reliable, authentic, and accurate, even if this is not based on the same legal or technical footing as public records. And there lies the rub.

Archives used for public records systems are normally devised by archivists or record-keeping specialists in such a way as to reasonably assure trustworthiness. In the paper world, a chain of custody is established by independent and redundant trusted parties. They attest to signatures (i.e., seals) that become part of the document as it moves from party to party for signature; take custody of the document and retain it in a secure location; track it in the fonds through numbering, ordering, cross referencing, and other related processes; indicate how, when, from whom, and other characteristics as documents are ingested, stored, moved, retrieved, transmitted, examined, copied, migrated, and so forth; and generally keep records of their activities which are transparent and made available for examination.

This all depends on trust in the custodian as somebody who has not altered the records and has not allowed anyone else to do so. This latter requirement, that of not allowing others to alter the records, is problematic in the Internet in general because it is not designed or built for this purpose.

Examination can detect inconsistency in and between records and fonds and this supports trusting (or challenging) the trustworthiness of the records.

But this is not the case for depictions presented by the Wayback machine. Collections are made on a seemingly arbitrary time frame from subsets of automatically selected Web sites. Different components that form a visualized Web page are collected at different times, stored with only a single reference to a collection date, and are not attributed or tracked in all of the other ways archives are managed. They are not systems of records as much as amateur collections, but they are sometimes treated as if they were traditional archives.

In the digital world, alteration can happen unintentionally or intentionally, the state of the art in protection of the WayBack Machine is not transparent, and its adequacy has not been established by a scientific or rigorous process. It does not apparently follow the rigors of archival science or records management, and thus it should be inherently obvious to an expert in the field that it does not have the same status as public records or archives maintaining and operating within those standards of care. This is also the case for many other Internet-based sites asserting archival or records status, and this is one of the important reasons a science needs to be developed in this regard and diplomatics must be developed as a field to question such documents.

The situation is further complicated by the fact that the mechanisms of the Wayback machine change over time, are not externally well documented or transparent, and do not follow widely accepted archival principles. In fact, once the findings discussed here were made public, the Wayback machine was changed with only minimal notice and little apparent transparency. Thus there isn't external repeatability across those changes, a basic foundation for scientific fields, and doing an accurate reconstruction becomes problematic.

Legally, depending on the bent of the judge and the precedence from cases that may reflect previous mechanisms or poorly tested assertions, depictions that are not accurate, reliable, or authentic, may ultimately be admitted, presumed trustworthy, and treated with a weight similar to that of records maintained by government bodies or real archives.

D. *Some related information on records*

While this situation may seem problematic, the reality of digital records is in general quite tenuous compared to other forms of records previously used. Some examples from personal correspondence may be informative, and some of them may be recognized as related to stories in the popular media.

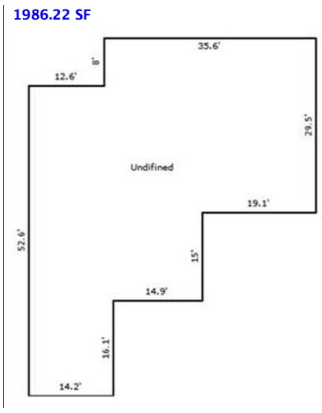
- A global non-government agency (NGA) indicated that in some cases, they hold records where 80% are of unknown type. When asked whether assistance was desired in trace typing them, the response was that, while they must maintain these records, they actually have no resources or desire to type them. Their obligation stops at proper retention.
- Migration of records from system to system over time is necessary for retaining the utility of these records because digital systems fail and older systems are no longer available, while newer systems don't support all of the mechanisms of the older systems.
  Conversion is thus part of migration of records, and the result is that migrated records are often, at best, an imitative copy, sometimes a simple copy, and sometimes a pseudo-original copy. They may never be viewed as they were initially formed, and loss of utility in such conversions is not uncommon. Part of the migration problem faced in digital archives includes creating the necessary mechanisms to be able to produce copies in one form or another of the records and identifying and recording the nature of any changes associated with conversions and non-original mechanisms in terms of what is then depicted and what is no longer depicted.
- In forensic archives of legal matters, there are often large volumes of data collected and retained that are in unusable form because they have not been migrated or converted and the original mechanisms and/or context may no longer exist to meaningfully reconstruct or operate them. Given that appeals processes may come many years later, this evidence may no longer be viable in those processes should retrial or re-examination be required.
- Many modern devices and systems are complex and lack transparency to the point where mechanisms of their operation are not reasonably discernible. Furthermore, the patch automation in place today often results in situations where exact versions are not available and may be difficult or impossible to accurately reconstruct. Thus, establishing causality in reconstruction may not be accurate. The field of reconstruction becomes very complex in this light.
- Governments have now admitted covert methods used to alter the seeming operation of mechanisms, thus making them act in ways unknown even to their manufacturers. This potentially shakes to the foundations the notion of keeping archives that accurately reflect the reality of what took place. The competition to rewrite history and current affairs in the digital realm would seem to present problems for the trustworthiness of digital records for legal purposes.
- Some nations and other similar entities now use exclusively digital records to reflect the operations of their governments, including without limit, the original writing and official codification of their laws and legislative history. This includes scanning documents that become part and parcel of the legal constructs of their societies, as well as born-digital records.
  Recent revelations identify that scanning devices no longer simply make representations of pixelated color values in digital form with known accuracy and precision limits. Rather, some of these devices now read the content of documents and rewrite them, sometimes replacing digits, words, spelling, and other elements of content with "corrected" versions. The very laws codified in statute and then used to make decisions about peoples' lives cannot be relied upon to accurately reflect the laws as passed, and things like financial records may not be accurately recorded for future analysis such as taxation and issuing fines.

It appears that there is a potentially desperate need for a questioned digital document science. This field might be reasonably called digital diplomatics, named after the existing diplomatics field built for the same purpose in the non-digital realm.

Part and parcel of diplomatics was the development of archival science and records management, and the same path would seem a reasonable trajectory for digital diplomatics leading to and helping to guide digital records management, digital archival science, digital records forensics,[21] and the broader field of digital forensics.

III. How were these cases resolved?

A. *Case 1*



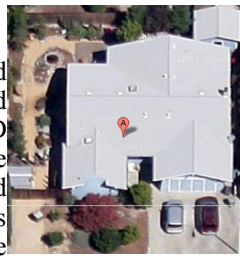
A building inspector's analysis of Case 1 yielded the drawing of the property given in Figure 1B. The resulting calculation of square footage was 1986.22 sq. ft. This is obviously at odds with the overhead imagery of the house seen in Figure 1C. The inspector was looking at the records of livable interior space in city permits, the full details of which were no longer available from the relevant time frames.

Figure 1B – The inspector's version of the house

21 L. Duranti, "From Digital Diplomatics to Digital Records Forensics", Archivaria 68, pp 39-66., 2009.

The overhead picture in Figure 1C shows a Google maps aerial of the house at about the time of sale. Note the substantial difference between the shape in the inspection report and the actual shape from the overhead. The dispute at this point was whether and to what extent a remodel of the former garage was properly accounted for in the calculation.

Figure 1C – An overhead picture of the house at issue



The issue was ultimately settled when county records were retrieved from the county archives. Figure 1D shows the official county report page used for taxation calculation and identifying that a laundry room was counted as livable space in the previous city remodel.

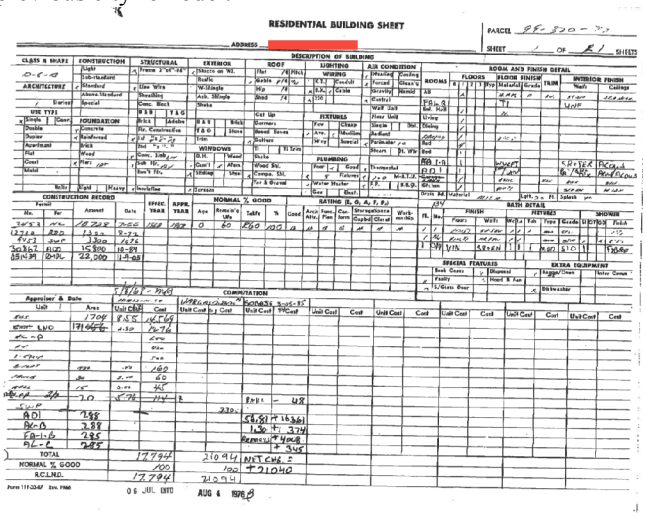


Figure 1D – County records from archives

When this final piece of the puzzle was introduced, the dispute rapidly settled with the sale square footage matching the original offering, the tax numbers, and the final sale size.

B. Case 2

Case 2 never made it to court and was settled prior to trial when both sides agreed that the digital records were inadequate to settle the dispute one way or another, and no other records could be demonstrated to resolve the issue more definitively.

Figure 2B shows the time line of appearances of different elements of the depicted site based only on the dates of the Wayback machine filenames associated with the different versions collected. In this case, the computers and content associated with the original activities were no longer available by the time the legal matter started, so no other provenance information was available. As such, the Wayback machine content was asserted to be the 'best evidence', and was in fact the only evidence supporting the asserted claim.

Figure 2C shows the time sequence in different terms. Note the dates and times are such that there is no date and time at which the second company (APM) can be definitively shown to have simultaneously appeared with the former company (RC).

It cannot be proven from this information that they did appear together, and it cannot be proven that they did not appear together.

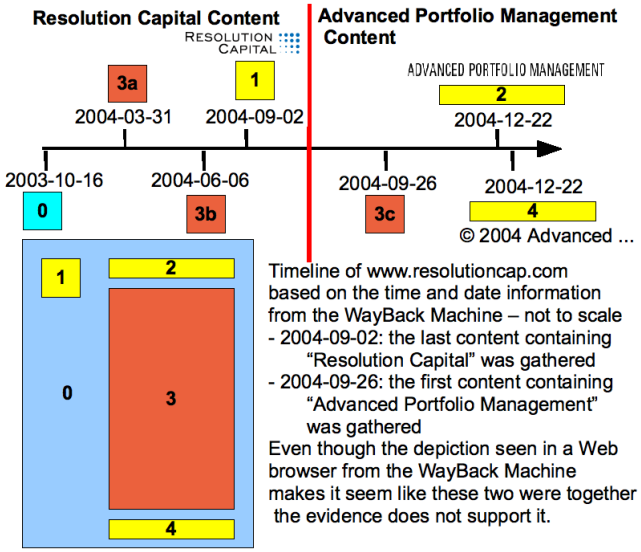


Figure 2B – The time sequence of the site

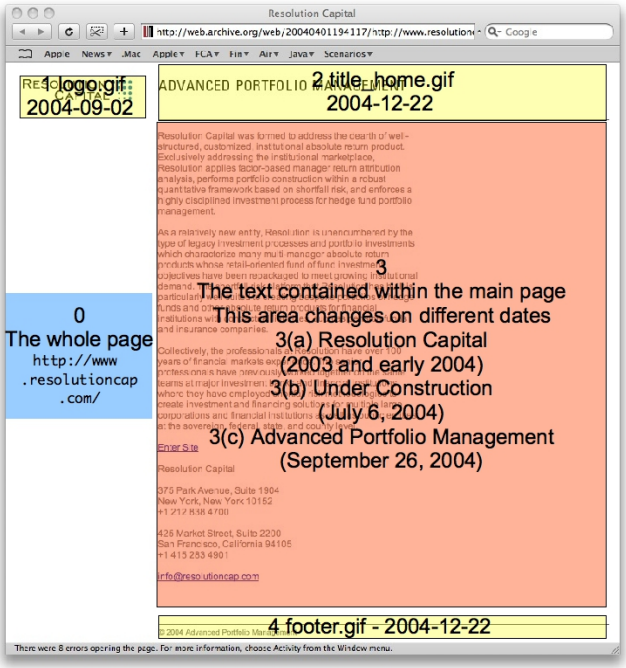


Figure 2C – Depiction areas with the histories detailed.

In this particular case, the screen images depicting the simultaneous appearance of both companies (Figure 2) is deceptive in that it appears to support a highly probative fact that is also highly prejudicial. But while it is certainly prejudicial, it is not actually probative, because it cannot be shown to be reliable.

The demonstration used to clarify this is depicted in Figure 2D. In this contemporaneous example, a Web site depicted as from 1997 on the Wayback machine was used to demonstrate the appearance of later content as if it were from

a prior time. The example substituted a graphical image instance shown in Figure 2D for an image not previously saved by the Wayback machine and thus depicted on the Wayback Machine as from the earlier date. The example was intended to demonstrate that either the Wayback machine depicted events as simultaneous when they were not or the author could predict the future (or time travel) including providing future pictures and facts normally not predictable.

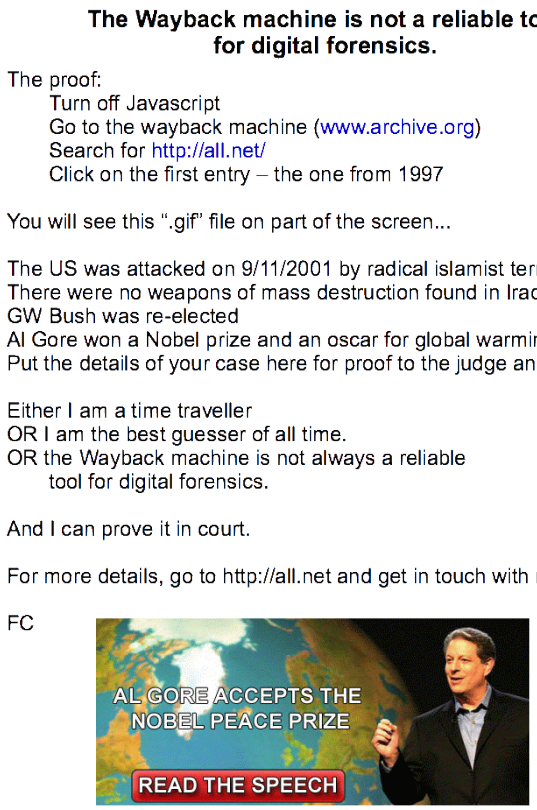


Figure 2D – Image used to demonstrate inconsistency

While this was an effective demonstration at the time, and was recorded as part of the report generation for the case at that time, the operation of the Wayback machine was subsequently changed as to not display such images under these conditions any more.

At this point, we are faced with a serious challenge for digital diplomatics. Since the Wayback machine no longer allows demonstrations of these sorts of failures to be easily generated and evidence collected for legal matters from prior dates may have these misleading depictions, there is no longer a reconstruction path readily available to demonstrate that false depictions gathered from before the change may be false. Rather, we are left with potentially probative and highly prejudicial digital traces and no way to demonstrate that they are not probative. At this point a best evidence argument along with a claim of "generally reliable as business records", or an archival ancient records claim could get such evidence admitted unless the digital diplomatics field becomes a part and parcel of digital forensics and such results are accepted in the relevant scientific community.

It may be reasonably shown that, for potential evidence gathered prior to the date of the changed operation, the demonstration done for Case 2 is adequate to question the document, and this may be used in conjunction with a more theoretical path including the more cogent argument about cause and effect in light of time lines. However, all of the theoretical points are likely less effective than a simple demonstration of predicting the future.

If the Wayback machine didn't use date and time stamps as pathnames and store them with reasonable accuracy in some portion of the instances involved, this approach would not work. Indeed, there is no real assurance that the time mechanism of the Wayback machine is generally reliable or reliable in any given case.

IV. Broader implications and a path forward

A. Implications

While we know about the Wayback machine, this is only one example of a potentially unlimited set of similar challenges faced in the digital evidence arena today. If traces are not properly collected along with the related information forming the archival bond and redundant data about the archives and their operation at the time the traces were identified and collected, by the time a legal matter gets to the point of examination, the information required to question the documents may be gone. The rapid changes of Internet sites combined with the lack of transparency, records of past versions, and retained audit information, and the proprietary nature of many such sites, makes reconstruction infeasible in many cases. And without such reconstruction, the seemingly probative information admitted as normal business records or under some other similar exception to hearsay, may prejudice such cases to the point where injustice is regularly done.

B. A path forward

It seems that one of the vital components contained in historic archives and systems of records is missing from the digital arena today. That is the various elements of records and record-keeping producing what is sometimes called metadata, context, provenance, chain of custody, and transparency (i.e., the archival bond). This includes a variety of different things that are becoming vital to addressing the discovery issues in digital evidence cases.

In addition, courts are hesitant to allow collection and analysis of entire systems and mechanisms because of minimization concerns (criminal) and costs (civil) associated with electronic discovery, and for very large systems (e.g., Google's gmail system), practicality prevents examination of the totality of the collection and fonds.

We will assume for the moment that tasks necessary for forensic examination are to be collected by a forensic professional (i.e., a diplomatist, examiner, or trained digital evidence collector) who is engaged by a party to the matter or an independent party with appropriate interests. What then might be reasonably collected and documented to assure proper diplomatic analysis?

While specific details for different circumstances remain elusive, some examples of the information reasonable and prudent to forensic use and diplomatic examination include, without limit:

- Date, time, and detailed actions of all activities performed by the collector, taken by them as contemporaneous notes at a suitable level of granularity. Who did what, with what tools, when, and what were the results. This should include the ability to reproduce results. So for example, if a command line is used, files should be kept and commands recorded with results, and relevant files referenced in the notes, or as part and parcel of the report as generated. In cases where repeatability is not feasible (e.g., real-time collection of network traffic), records should include details of dropped packets and other similar information as available, and to the extent feasible, redundant records from related mechanisms (e.g., network flow logs from routing equipment during the times of collection).
- The documentary forms as observed by users in the known various circumstances, including sample documentary forms from all potentially relevant presentations. These should be in imitative copy form that can be reliably viewed in as near to an identical fashion as the original, but which is entirely contained in the stored form without need to reference or display external content.
- All URLs, sources for all Web pages or other content retrieved and observed from systems over which the observer does not have direct control, and depictions in an imitative copy form.
- A copy of whatever can be reasonably attained in a computer usable form. For example, in addition to an imitative copy of a spreadsheet as depicted on a screen, the actual spreadsheet should be saved in as close to a copy in the form of original as is feasible.
- Records of an archive within which the information is stored should be retained to the extent feasible. Ideally, the process would use a transaction system that retains the history of all transactions, but alternatives such as periodic backups with the ability to go back in time for retrieval may be sufficient in some cases. Note that this is potentially problematic with discovery rules.[22]
- Elements of the archival bond, such as the directory information about storage locations, relationships among records and files, classification codes[23], sequence numbers, date and times associated with documents, etc. As an example, many practitioners retain files in dated directories with dated filenames, such as 2013-11-25 for files received on that date, versions from that day selected for retention, records of retrieved files, etc. Alternatively, sometimes these filenames are used for the date of the content (e.g., a paper published on a date might be names starting with YYYY-MM-DD- followed by other elements of the name.
- Other records from the systems used for the examination process. This includes test results for tools, calibration information for measurement mechanisms, records of activities performed with tools (e.g., records of commands issued to clear a disk before copying content to it), log files retained by the systems used in normal use, and other similar related data.
- Transparency information, such as copies of online contracts contained within the Web sites used in any retrieval process, details of how mechanisms work, documents from relevant manuals and related documentary sources used, and generally, all considered and/or referenced materials.[24]
- Supporting documents for named protocols, methods, tools, programs, etc. For example, when referencing the use of an Internet Protocol (IP) address or a Universal Resource Locator (URL), on first use, the relevant Requests for Comments (RFCs) should also be collected both for clarity and for historical reference and reuse. By example, we might cite http://www.ietf.org/rfc/rfc791.txt, which details IP version 4, as included with the report in a file named rfc791.txt in the Considered directory.
- Version numbers for everything identifiable, including major and minor versions, date and time stamps, and related indicators are often useful in settling disputes, but are also often unnecessary to the purpose, particularly in clear context.

As suggestions, these may be within the range of reasonable and prudent acts, but there remains the problem that they are only that. They are not widely accepted by the digital forensics or digital diplomatics community, are not comprehensive, do not provide substantial details of a suitable documentary form, are not structured so as to provide meaningful automatic use, and if and to the extent they are missing, they do not imply that the traces offered as evidence will not be reliable, authentic, and accurate, or will not be admitted, useful, reasonable, and appropriate.

Unlike the records management profession, which often has the opportunity to manage records from the "womb to the tomb", the archival and digital forensics communities must usually work with only the residue (archival) or traces (forensics) available. But when experts collect evidence (forensics) or participate in records creation (archival), it would seem useful to provide guidance and a standard approach as to what to collect and retain and what not to.

22 Discovery rules may require that drafts be retained and presented for discovery, which often creates more problems than it solves. This may be why the US Federal rules of civil procedure were amended to eliminate discovery of drafts.

23 The class of records in a fonds hierarchically organized in primary, secondary, etc. classes.

24 US Federal rules of civil procedure require retention and discovery of considered material as part of expert reports, however, many reports fail to contain substantial references.

It is important to recognize that the examiner gets what they get. While in many cases there are opportunities for discovery, in other cases there are not. Civil matters often involve parties who are uncooperative and cannot be forced to act against interest. Criminal matters have similar limitations associated with the right to not self-incriminate.

The natural course of events do not result in preservation at the point of inception, and as a result of lack of discipline by those who implement information technology, this leads to situations where certainty is hard to attain. Current metrics don't provide insight into the resulting certainty of analysis and this limits the realistic ability to place likelihoods on outcomes of examinations.

The best we can currently do is to identify consistency or inconsistency with hypothesized causes and mechanisms based on available traces and experience. The absence of evidence is not evidence of absence.[25] When no definitive answer exists, we must learn to say so, and as a community, we need to develop the methods of digital diplomatics and records management in order to give a reasonable hope of justice being determinable in disputes.

A path forward suggests the notion of applying the same criteria used for the inherent presumed trustworthiness of public records. In this approach, the independence and due care charges of public officials combined with redundant methods starting at the initiation of a public record are the basis for trust in the system. But carrying this to the full spectrum of potential traces that may be introduced in the legal system implies forcing criteria on the private sector that they may be unwilling to accept, and perhaps justifiably so. Perhaps the creation of a standard for assured admissibility would be a motivating factor, but this sort of approach has rarely succeeded in the past except for those who already have legal requirements for diligence.

At a minimum, those entrusted with the retention of public records should create and/or require the mechanisms necessary to provide the same level of certainty with respect to born-digital public records as for born-analog public records. The notion of public records and archives based on operation in the cloud-based computing environments of today seem, at first glance, to be oxymoronic. However, it may be reasonable to leverage the low cost and high performance of many public cloud-based computing environments for limited purposes, such as widespread rapid access without the same level of surety required for use in the legal context. A more thorough process may then be used for the official versions of records, which may almost always be identical to the unofficial versions, thus providing a combination of high surety when needed and accessibility.

V. Conclusions

We started out by identifying two very different cases involving very different facts, issues, and component parts. The commonality is that they both depend on documentary evidence demonstrating questioned document challenges. The difference was that one set of documents are born-analog and the other born-digital.

From a diplomatics perspective, born-analog documents are better handled because we know more about how to manage them historically, and they reside in a context that has been well worked out over centuries. They involve known causal mechanisms that can be reproduced and examined using stable scientific methods and principles with measurable levels of accuracy.

The born-digital documents demonstrated many of the problems faced in the current context, and the discussion identified many of the challenges we face in digital diplomatics today. Perhaps the underlying lack of an adequate scientific examination basis is a major part of the challenge we face, but there is also a major challenge in the manner in which records and other documents are generated, cared for, and produced.

The origination problem is particularly disturbing. The lack of a single identifiable documentary form that persists over the lifetime of the record seems to be part of the underlying problem that cannot be solved in the current paradigms of digital systems. While paper documents such as building permit records are altered over time as they are updated to reflect new information, digital documents, including modern building permit records that don't include an original paper signed documentary form, don't have the rich set of residues to examine. Instead, we have a collection of potentially distributed digital record components and other bit sequences associated with the fonds, many elements of which are not currently retained across migrations, and without the transparency or consistency across record-keeping systems required to examine in a common structured way.

We suggested an initial set of objective information that would be helpful in the collection and analysis of digital traces, but offer little hope of attaining all of this information when traces are provided by others. The examiner's role in this situation is often limited, and there is little to be done about it today.

Born-digital documents have a long way to go. From their inception through their attempted use in court, there is a need for improvement in the data used to support the traces found. Consistency analysis holds hope, but there is often too little data to allow determinations of external consistency, and the process is fundamentally one of refutation rather than demonstration of adequacy. Without some level of guidance as to adequacy, examiners are left with an unlimited open ended challenge of building up enough threads to weave together a cloth that opposition experts cannot tear asunder.

We suggest the notion of building toward a standard of adequacy based on the historical diplomatics discipline and its application in forming the concepts of archival science and the basis for trust in public records. In particular, applying the elements of independent actors responsible only to proper record-keeping and with no foreknowledge of any particular case acting in a reasonable and prudent manner with adequate redundancy against accidental failures to assure that records are reliable, authentic, and accurate, seems like a good starting point.

25 This is a well known folk saying in forensics.