
LIBRARY
BRITISH

Retrospective Science and Scholarship
Some Perspectives and Tools from Phylogenetics and Digital Forensics

Dr Jeremy Leighton John
Curator of eMANUSCRIPTS

Digital Research & Curator Team
DEPARTMENT OF DIGITAL SCHOLARSHIP

Friday 12 April 2013: Sequence Alignment

at the

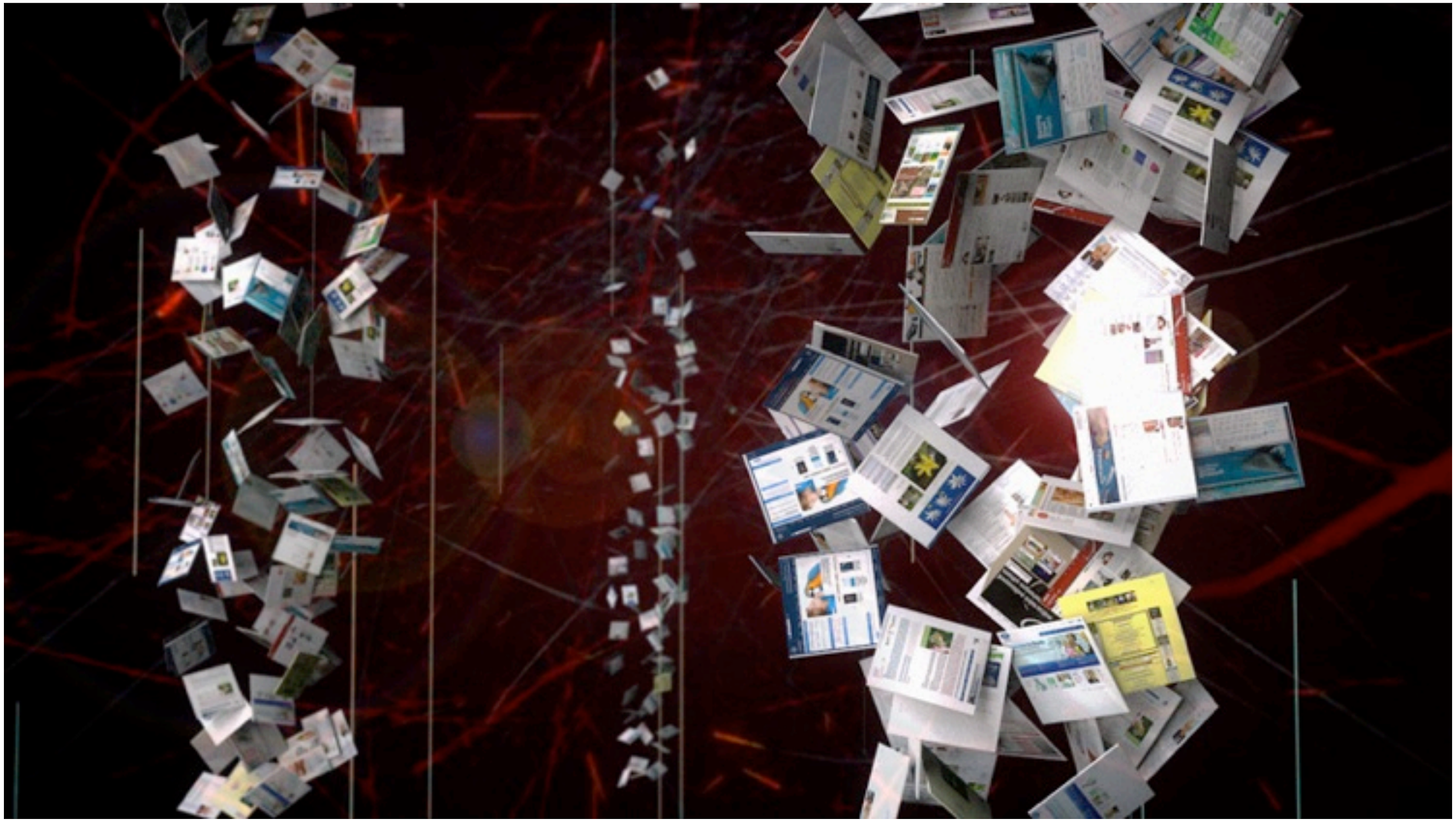
Shared Horizons: Data, Biomedicine, and the Digital Humanities Symposium

Maryland Institute for Technology in the Humanities: MITH
10-12 APRIL 2013

Cafritz Theatre, Clarice Smith Performing Arts Center
University of Maryland, College Park, Maryland

Click to Save the Nation's Digital Memory

Capturing the Digital Universe



The Legal Deposit Libraries Act established in 2003 the principle that legal deposit needed to evolve to reflect the massive shift to digital forms of publishing

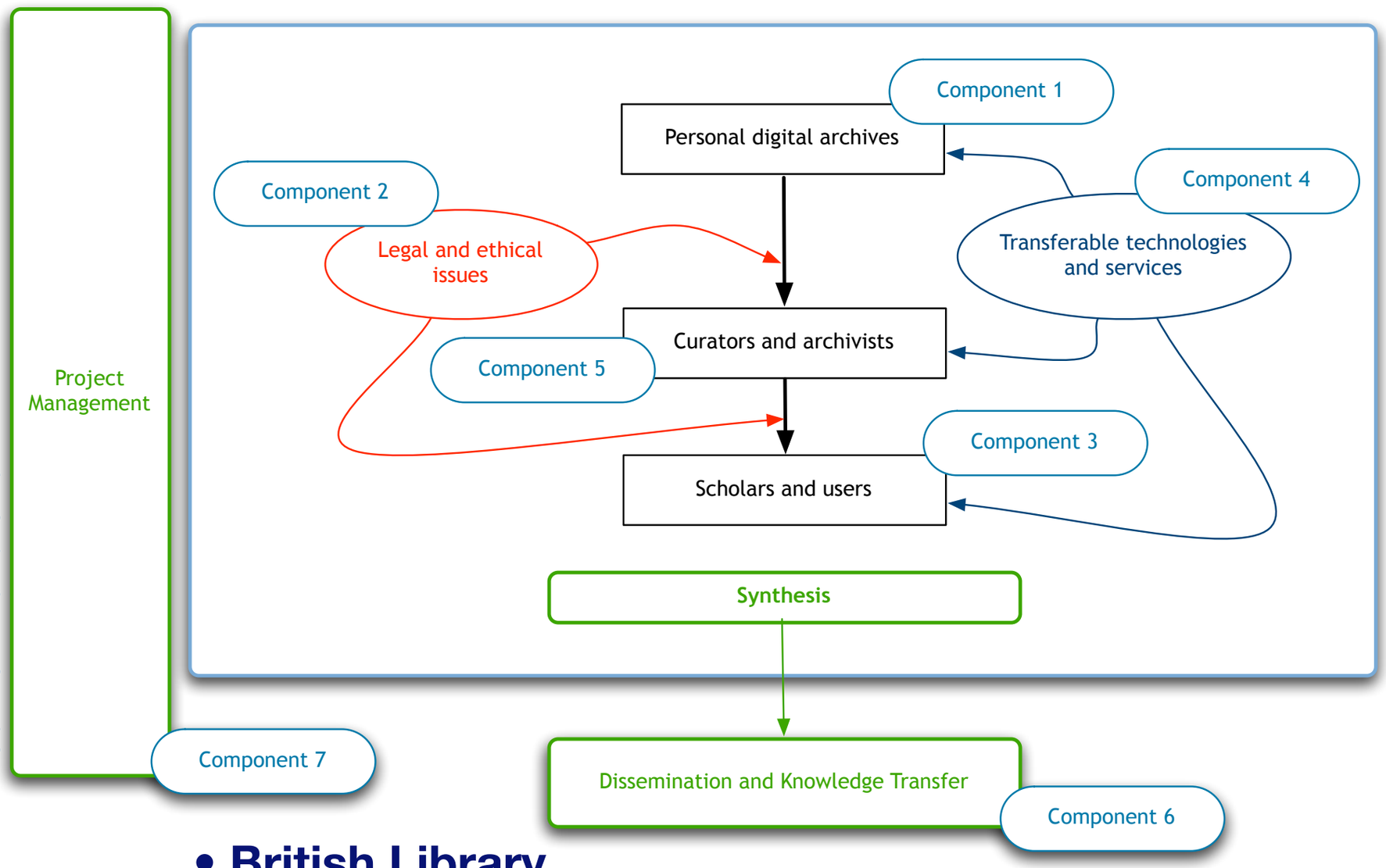
The regulations now coming into force make digital legal deposit a reality

Background

BRITISH LIBRARY

BRITISH LIBRARY DIGITAL >>LIVES

Digital Lives Research Project



graphic: digital lives: jeremy leighton john

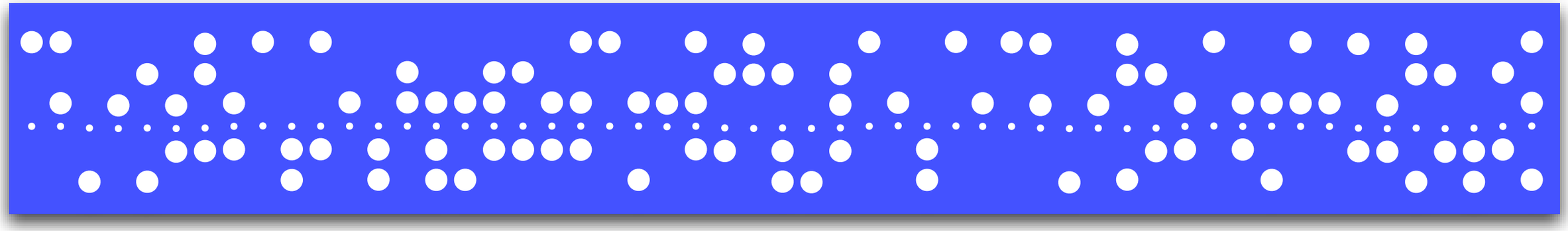
- **British Library**
- **University College London DIS/SLAIS**
- **University of Bristol CITL**

Arts & Humanities Research Council

BRITISH LIBRARY

Personal

THE DIGITAL MANUSCRIPTS PROJECT AT THE BRITISH LIBRARY



The eMSS Lab *at the British Library*

- authenticated capture
- audit control & reporting
- cryptographic and fuzzy hashing
- digital object export
- compound file expansion
- metadata extraction
- virtual media & portable forensics
- mobile forensics
- remote acquisition

Digital Forensics

Curatorial Examination

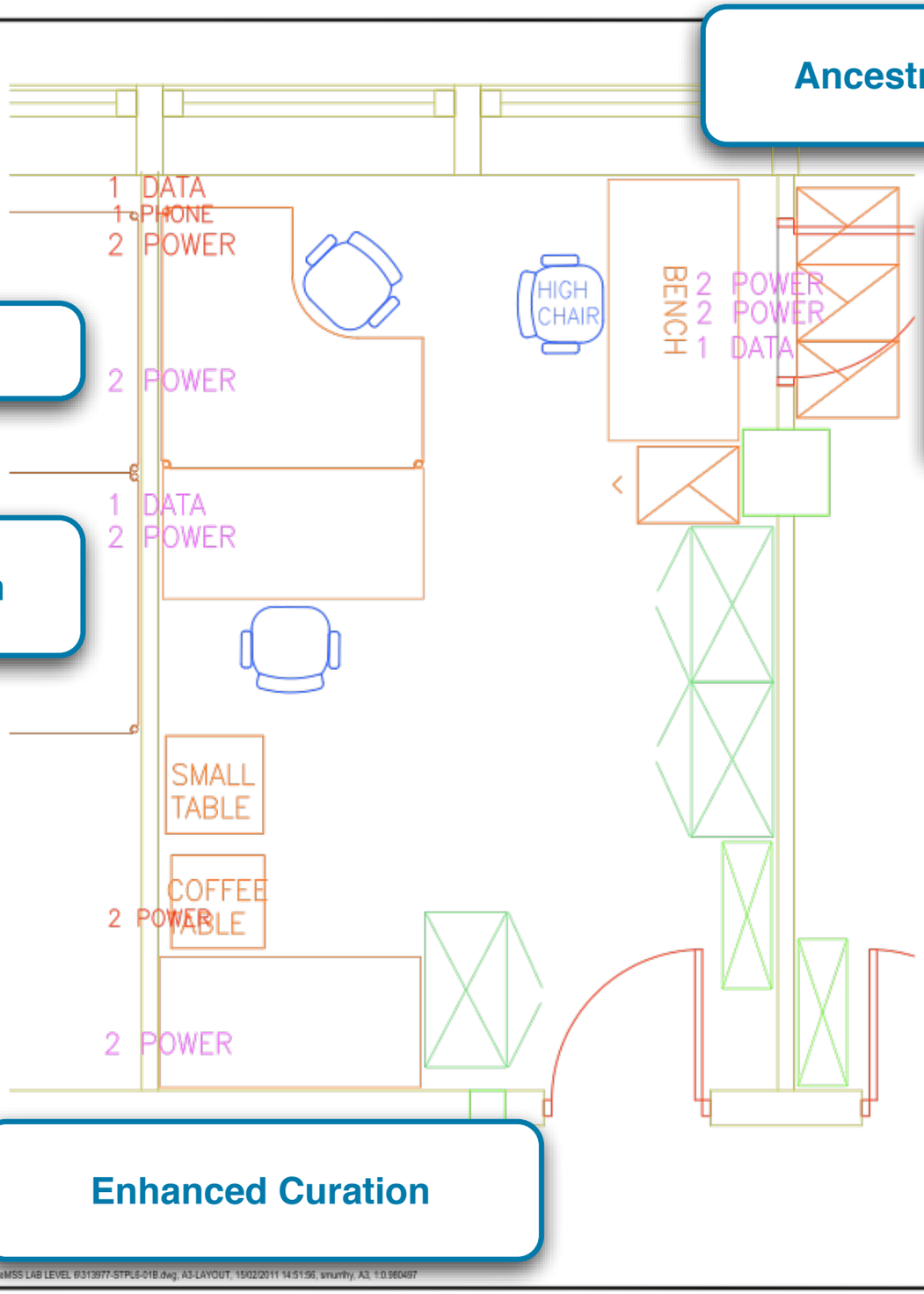
- inventory on reception (digital intake)
- metadata creation
- detailed description
- quality control
- digital rights compliance & due diligence

- panoramic stitching & immersion
- photographic enhancement
- 2D/3D graphics & visualisation
- video capture and editing
- fast digitisation
- community participation & social media

Enhanced Curation

Ancestral Computing

- hardware examination
- removal of internal media
- capture of ancestral media
- emulation
- virtual archival computing
- examination by originator & family

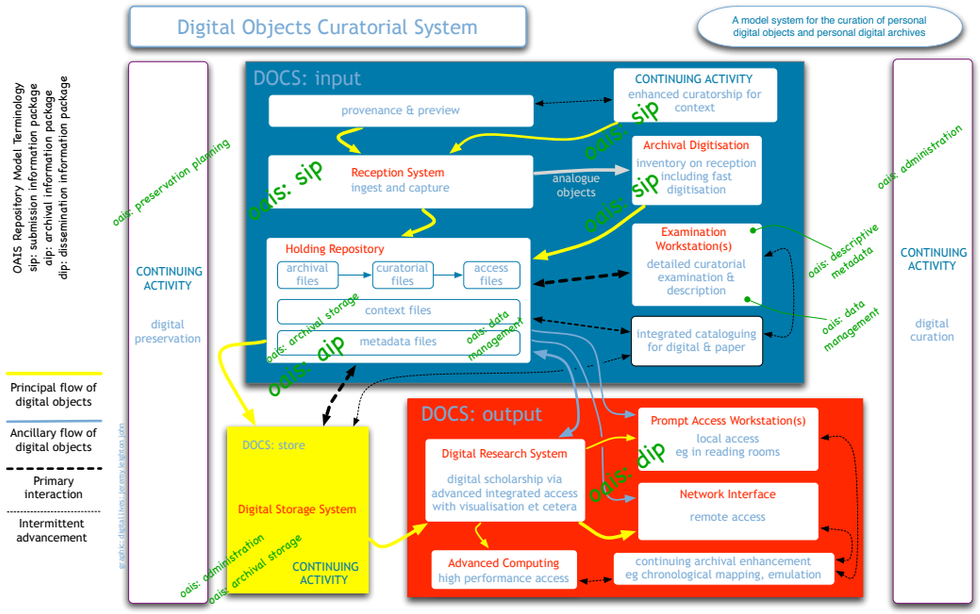


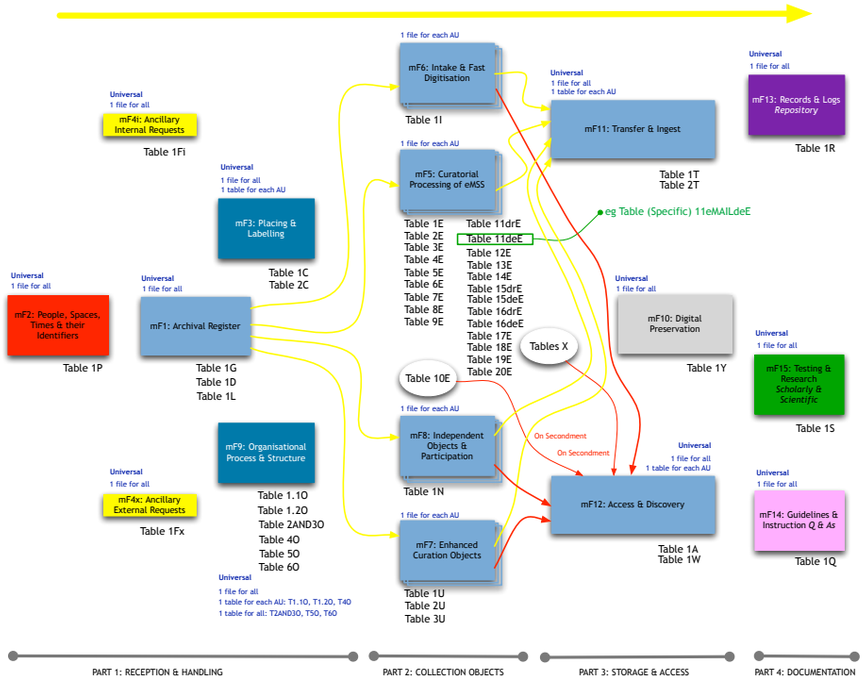
THE BRITISH LIBRARY
ESTATES & FACILITIES

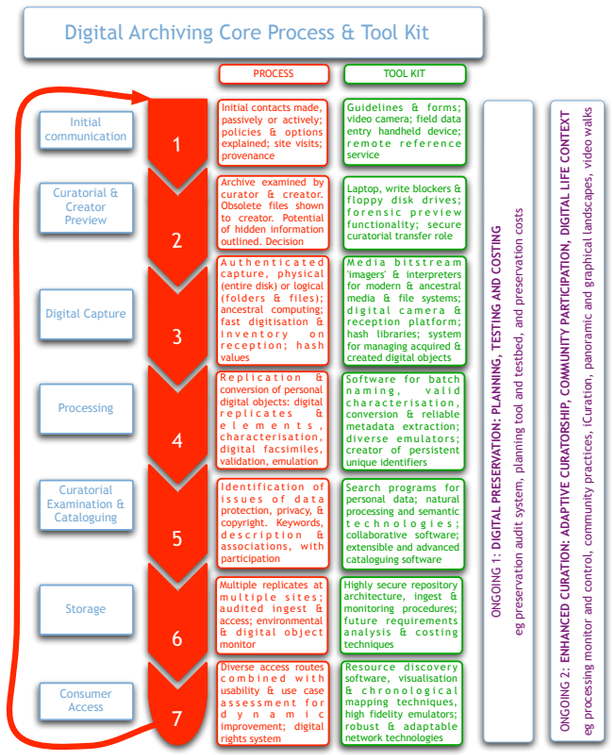
ENGINEERING SERVICES

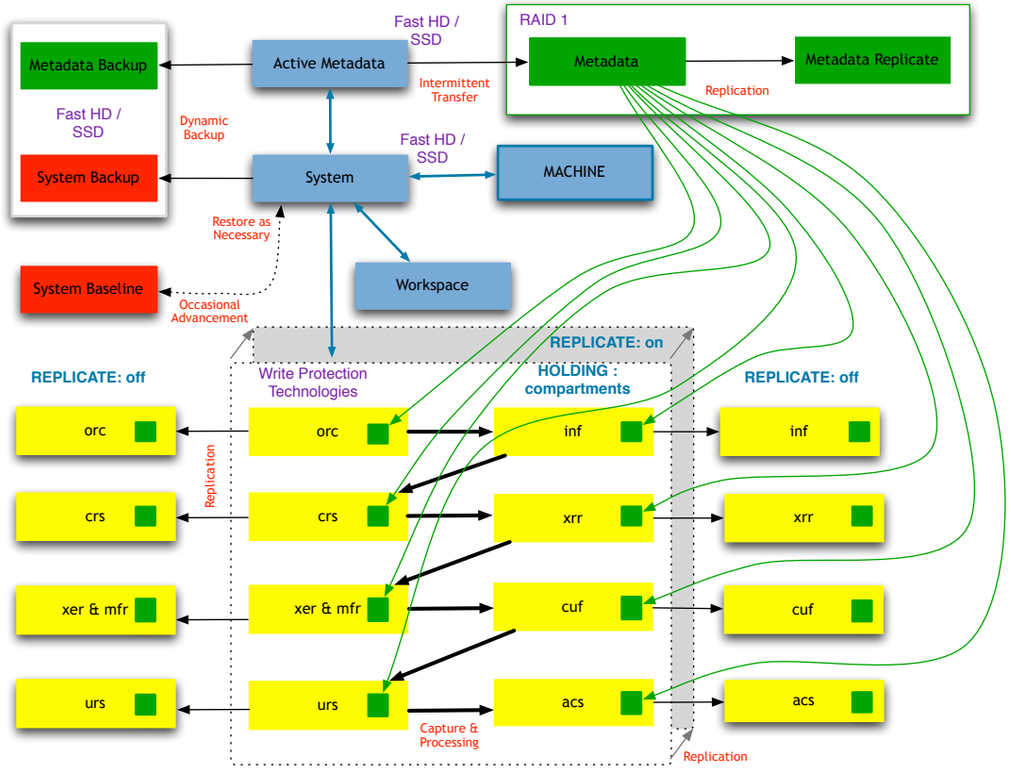
PROJECT
EMSS LAB SET-UP
ROOM 05

BRITISH LIBRARY







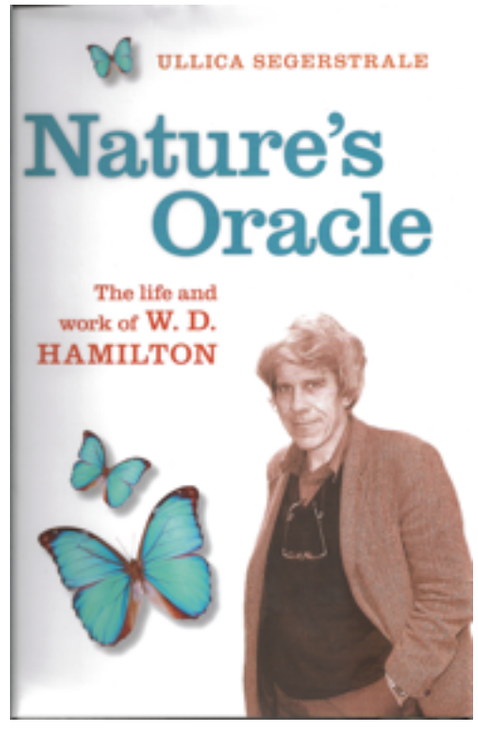


DOCS: DIGITAL OBJECTS CURATORIAL SYSTEM

WILLIAM D. HAMILTON

"…because topics in science often fade for no good reasons – certainly not because of error – it may sometimes help inspiration to be reminded of forgotten facts or to see new ones from an old point of view"

W. D. Hamilton
Narrow Roads of Gene Land, Volume 1



W. D. Hamilton Archive

JOHN MAYNARD SMITH

BRITISH LIBRARY

Manuscript curatorship

- authenticity and provenance
- confidentiality (including database protection)
- diverse physical nature: papyrus, wax on wood, parchment, vellum, paper
- associated artefacts & containers

Manuscript scholarship

- textual content
- styles, layout & decoration
- earlier drafts and amendments
- techniques and technologies of writing
- original objects, physicality
- fragments
- original behaviour

Digital Forensics

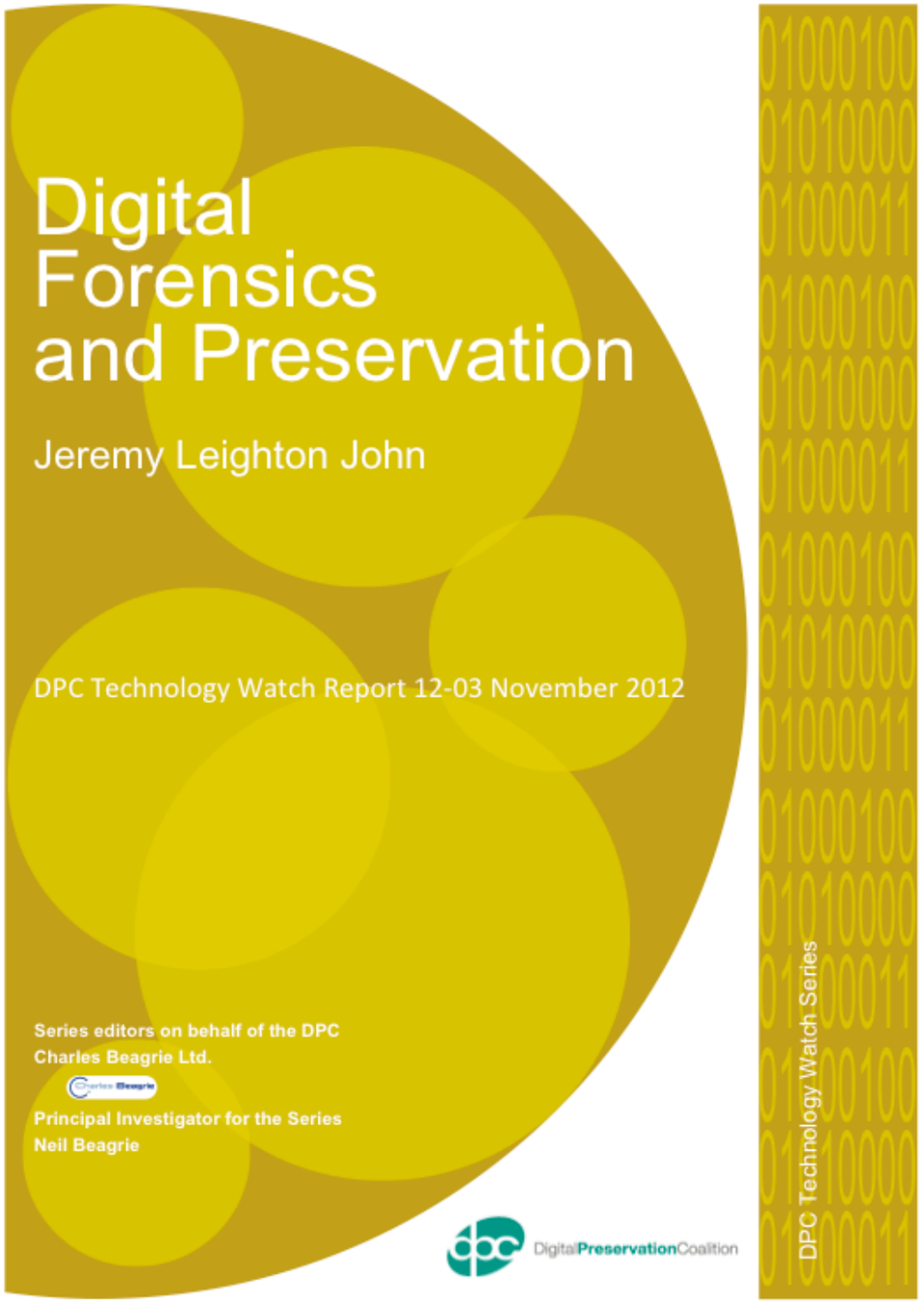
Digital Preservation & Digital Forensics

Digital Forensics and Born-Digital Content in Cultural Heritage Collections

by Matthew G. Kirschenbaum
Richard Ovenden
Gabriela Redwine
with research assistance from Rachel Donahue

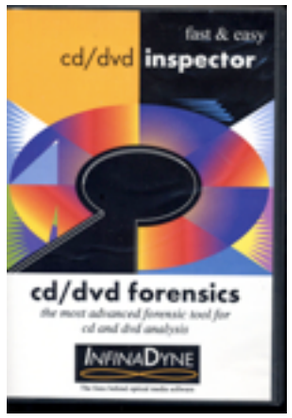
December 2010

Council on Library and Information Resources
Washington, D.C.



BRITISH LIBRARY

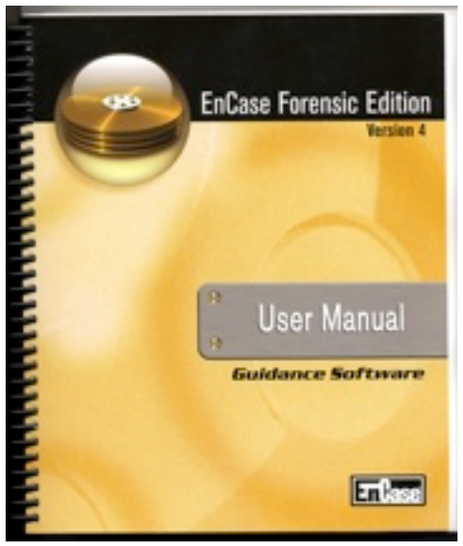
AFF®
Advanced Forensic Format



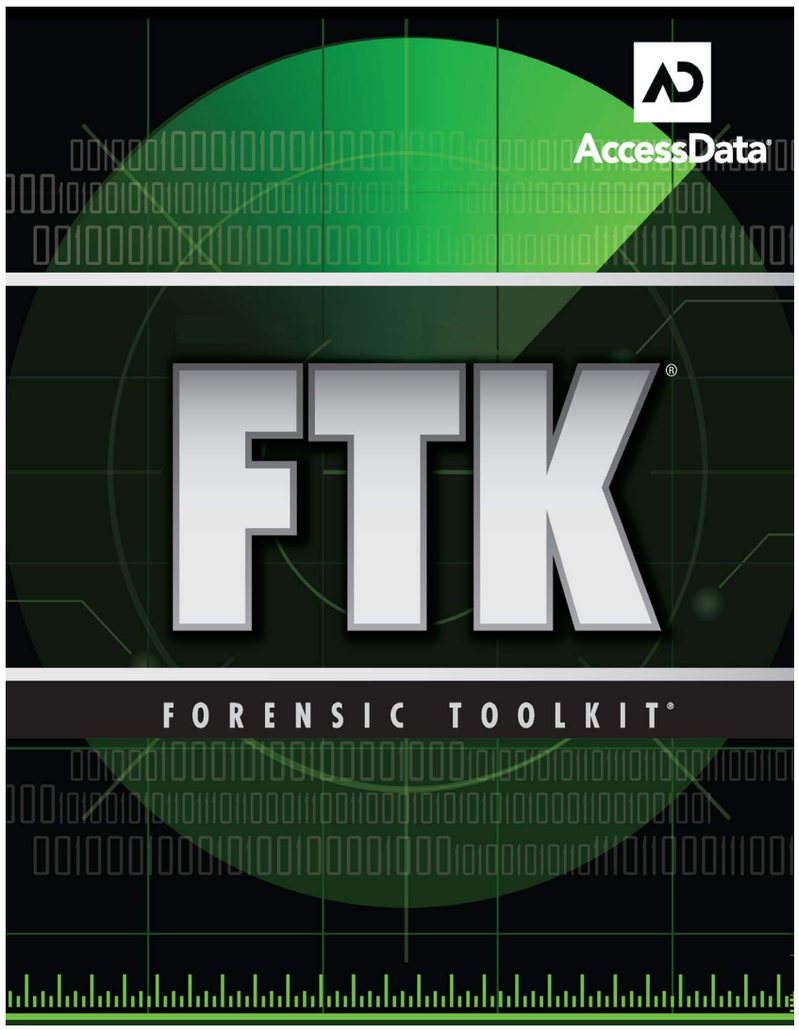
FORENSIC SOFTWARE

X-Ways Software Technology AG

X-Ways Forensics/ WinHex



MacForensicsLab 3.0 Manual



paraben's
e-mail examiner

The Sleuth Kit & Autopsy

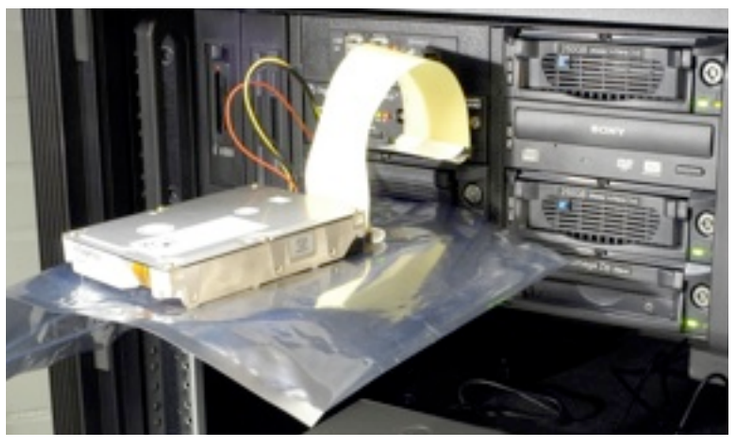
Home | Projects | Informer | Wiki | Support | About | Contact

sleuthkit.org is the official web site for The Sleuth Kit and Autopsy Browser. Both are open source digital investigation tools (a.k.a. digital forensic tools) that run on Windows and Unix systems (such as Linux, OS X, Cygwin, FreeBSD, OpenBSD, and Solaris). They can be used to analyze NTFS, FAT, HFS+, Ext2, Ext3, UFS1, and UFS2 file systems and several volume system types.

The Sleuth Kit (TSK) is a C library and a collection of command line tools. Autopsy is a graphical interface to TSK. TSK can be integrated into automated forensics systems in many ways, including as a C library and by using the SQLite database that it can can create.

Modern Forensics on a Mac

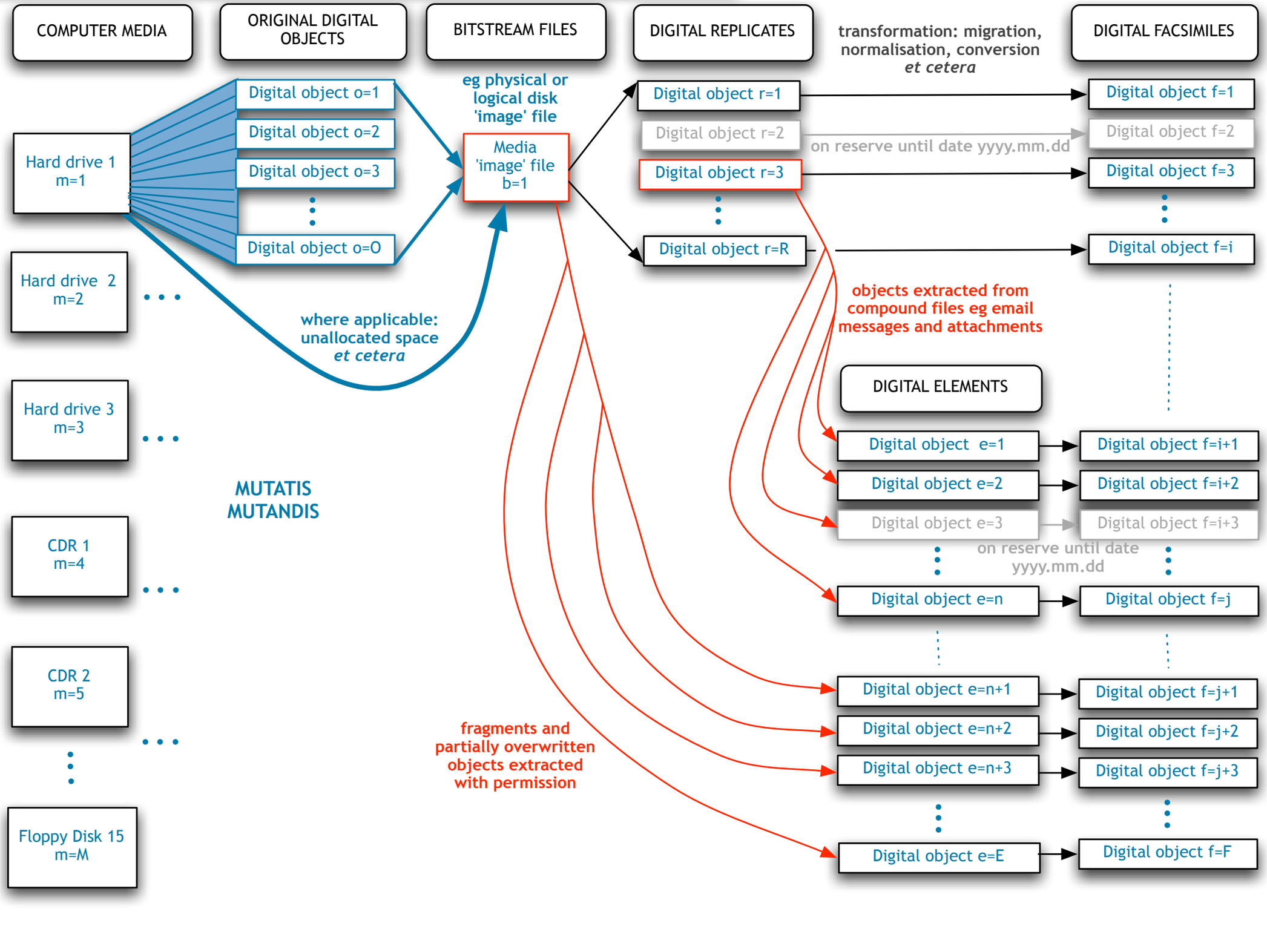
DIGITAL CAPTURE WITH WRITE BLOCKERS



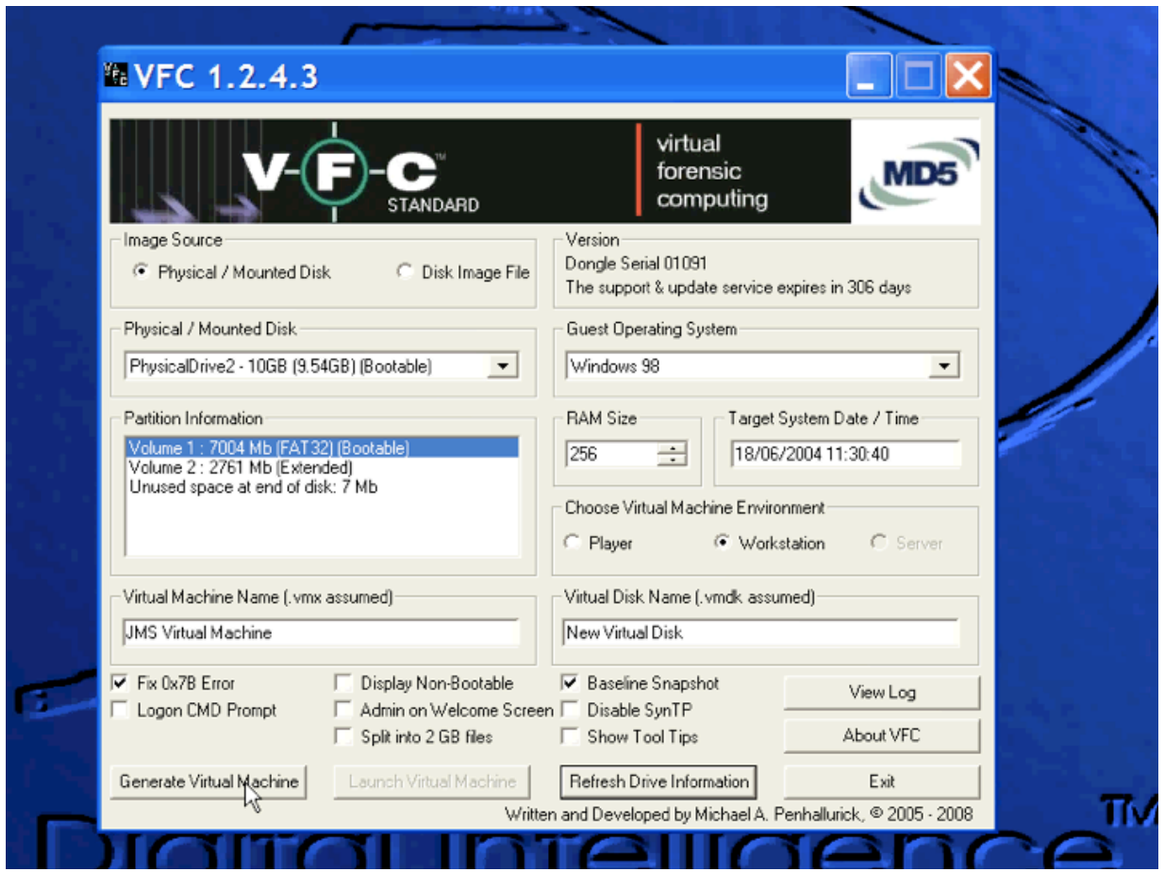
Forensically sound bit-for-bit replication of write-protected original hard disks, with verification (two hardware/software systems for testing for same hash values from a single collection hard disk)

DIGITAL OBJECT EXPORT & EXPANSION

Origins of eMSS: Personal Digital Objects



VIRTUAL ARCHIVAL COMPUTING & DIGITAL MATERIALITY

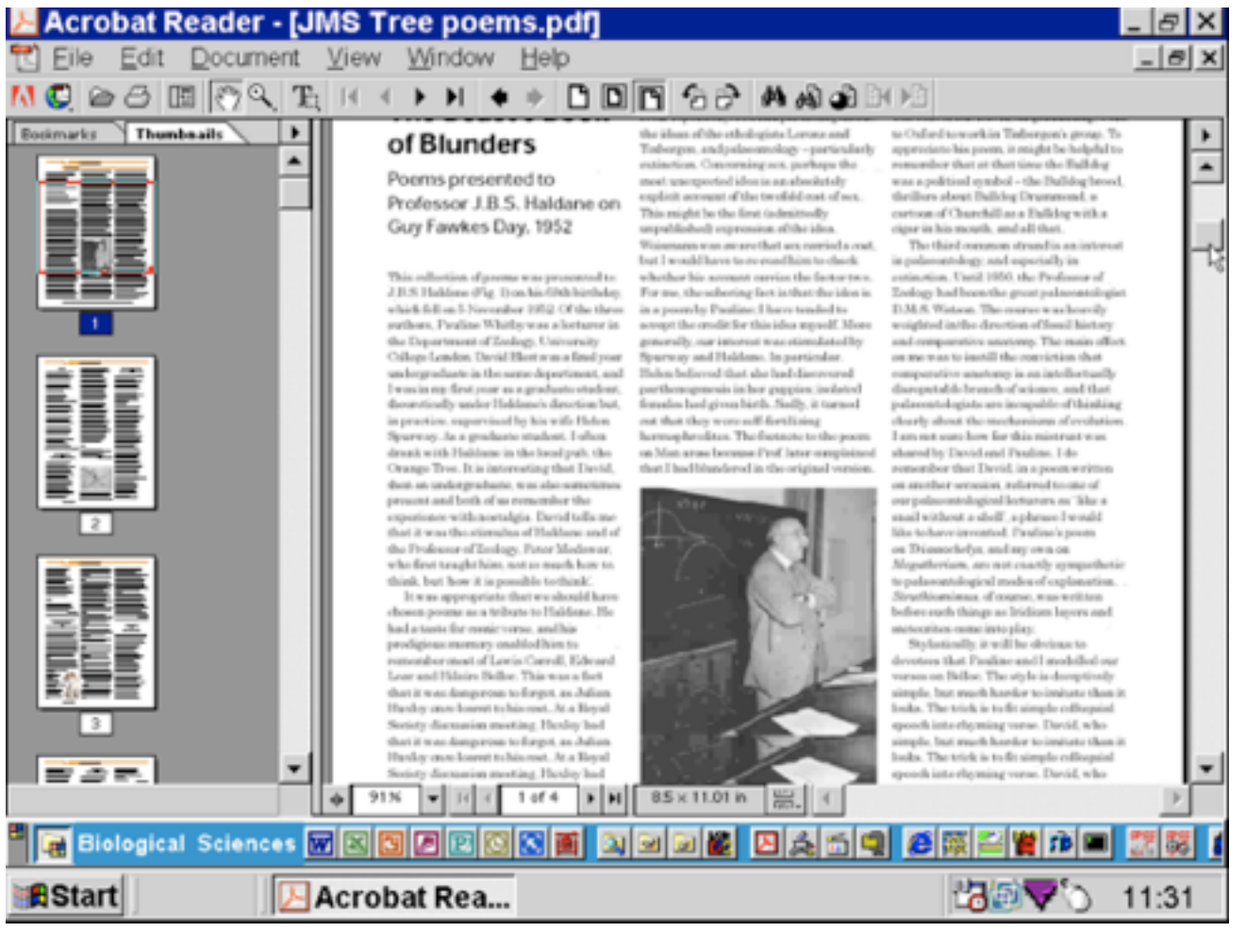


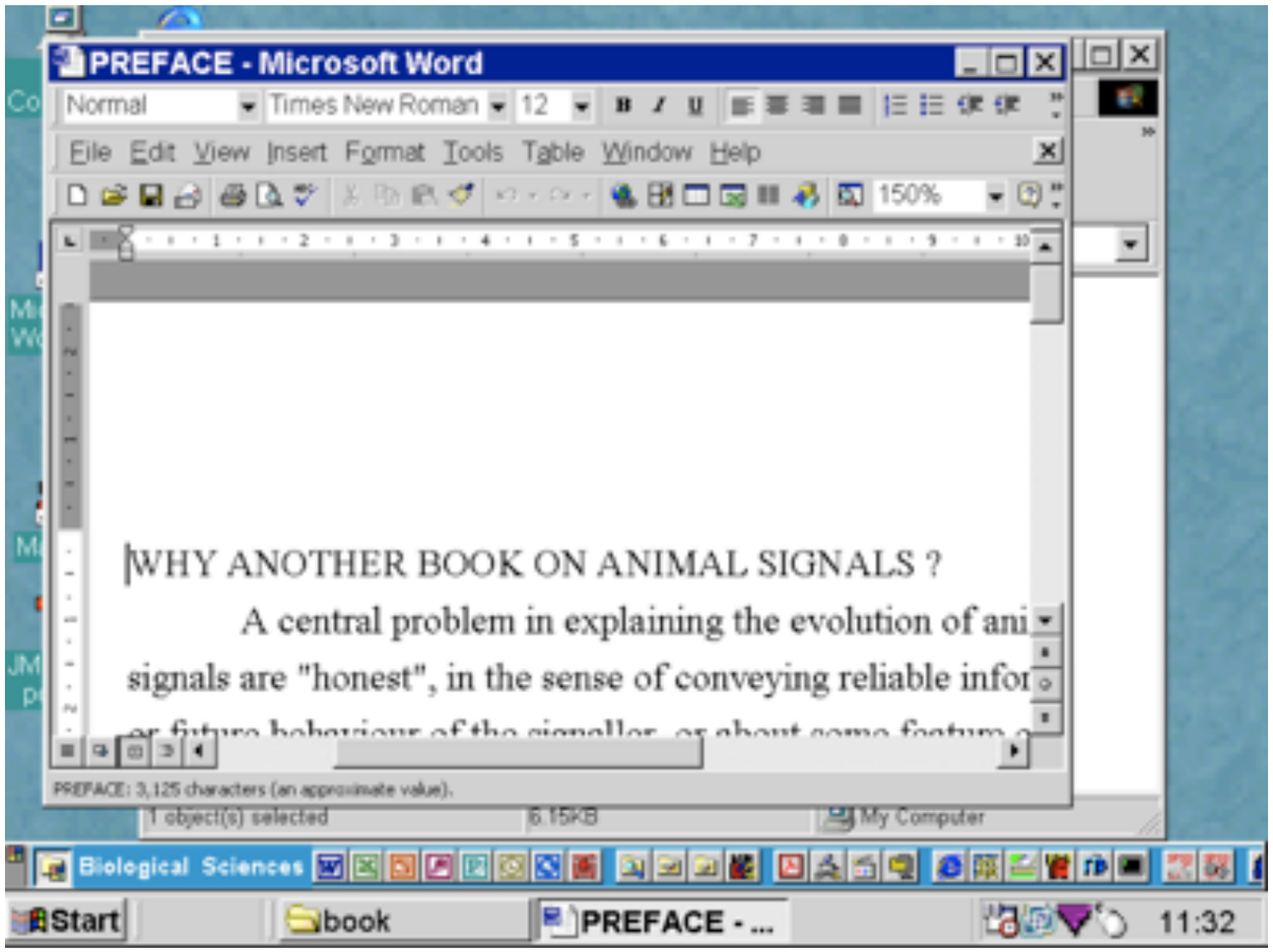
John Maynard Smith Archive

JMS extracted Internal Hard drive with Microsoft Windows


Distributed by:
Viglen Ltd
Microsoft
Microsoft
Windows98


My Computer
Internet Explor...
Microsoft Word (2)
Recycle Bin
School of Biological Sciences
Maple 6
MS-DOS Prompt
JMS Tree poems
Biological Sciences
Start
11:31


Acrobat Reader - [JMS Tree poems.pdf]
File Edit Document View Window Help
Bookmarks
Thumbnails
of Blunders
Poems presented to Professor J.B.S. Haldane on Guy Fawkes Day, 1952
Biological Sciences
Start
Acrobat Rea...
11:31


PREFACE - Microsoft Word
File Edit View Insert Format Tools Table Window Help
WHY ANOTHER BOOK ON ANIMAL SIGNALS ?
A central problem in explaining the evolution of ani
signals are "honest", in the sense of conveying reliable infor
PREFACE: 3,125 characters (an approximate value).
Biological Sciences
Start
book
PREFACE - ...
11:32

BitCurator Tools for Digital Forensics Methods and Workflows in Real-World Collecting Institutions

The BitCurator project is a joint effort led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) and the Maryland Institute for Technology in the Humanities (MITH)

The BitCurator project is collaborating with the Open Planets Foundation (OPF) which originated from an international research project led by the British Library

Open Planets Foundation

A community hub for digital preservation

BRITISH LIBRARY

Retrospection

Digital Forensics has two core aspects:

(1) digital capture of evidence and the protection of authenticity

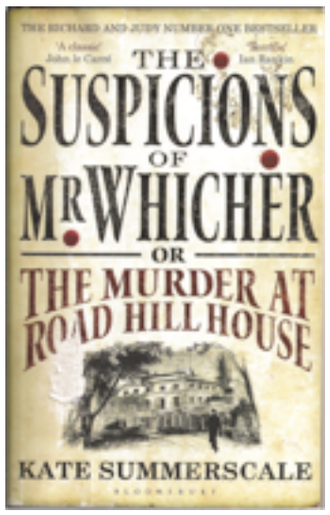
(2) analysing and reconstructing past events

History of Forensics

In 1842 Scotland Yard was formed, with a contingent of eight detectives

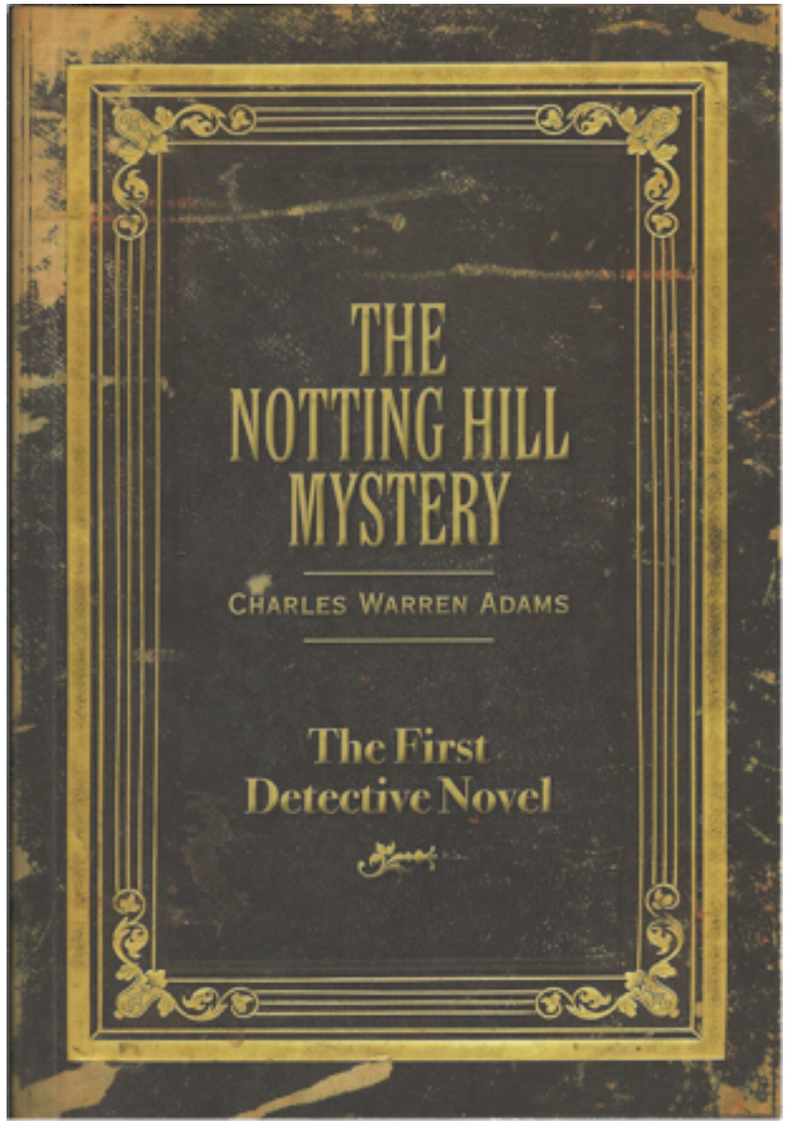
An account of an investigation conducted by one of them, Jonathan Whicher, some years later, beginning in 1860, illustrates the emergence of detective procedures

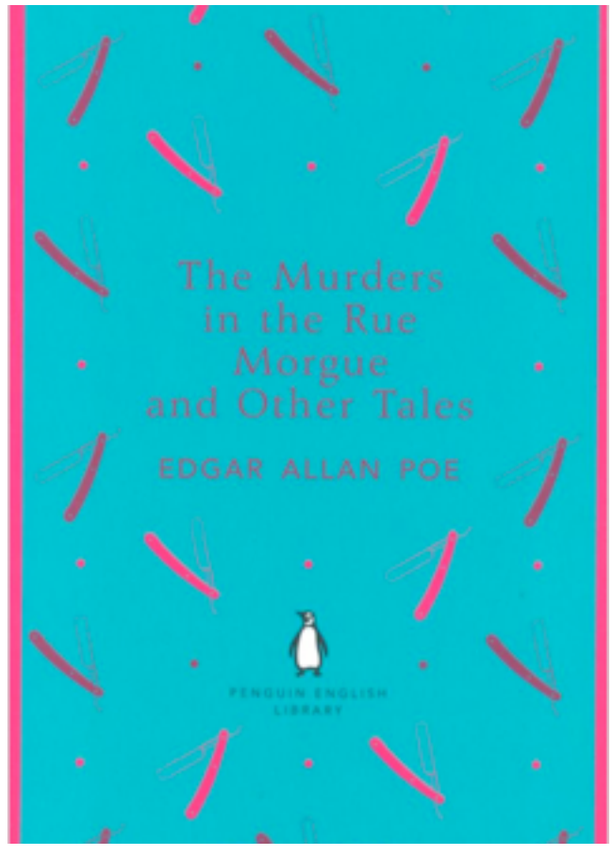
“The detective’s job was to reconstruct history from tiny indicators, clues, fossils.” “Like the natural historians and archaeologists of the mid-nineteenth century, Whicher tried to find a story to bind the fragments he had found”

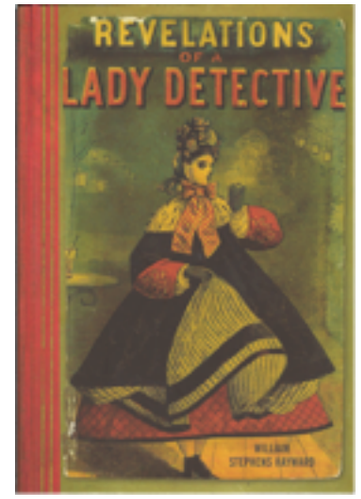


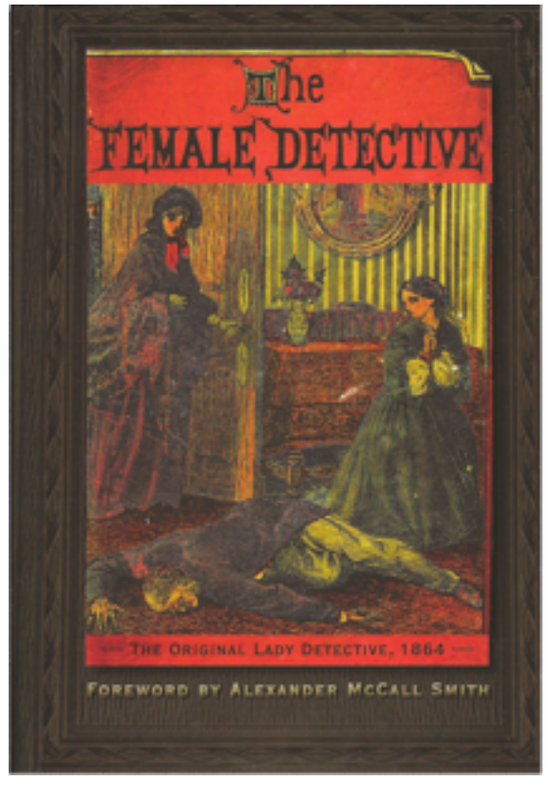
Kate Summerscale (2008)

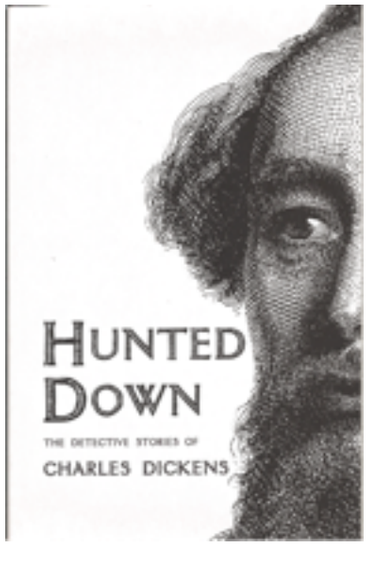
History of Forensics











Edgar Allan Poe and other early detective literature preceded Conan Doyle but it was Sherlock Holmes who seems to have most captured the imagination of investigators and detectives

History of Forensics

Image from Wikipedia

Alexandre Lacassagne

Crédit photo: BM Lyon fonds Vermard

Edmond Locard

“One must know how to doubt”

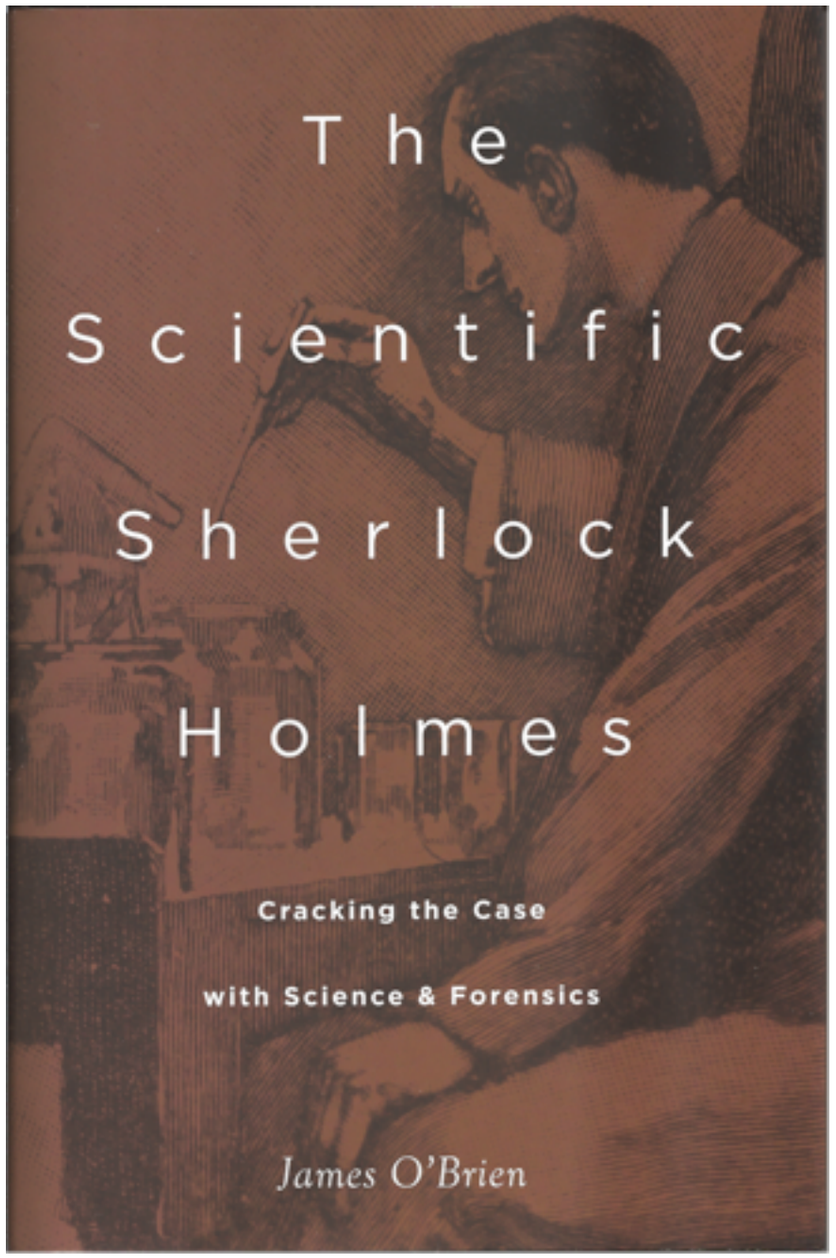
History of Forensics

Douglas Starr (2011) The Killer of Little Shepherds

Lacassagne and his student Jean-Henri Bercher published reviews comparing the methods of forensic scientists with those of Holmes

Shared methodology included:

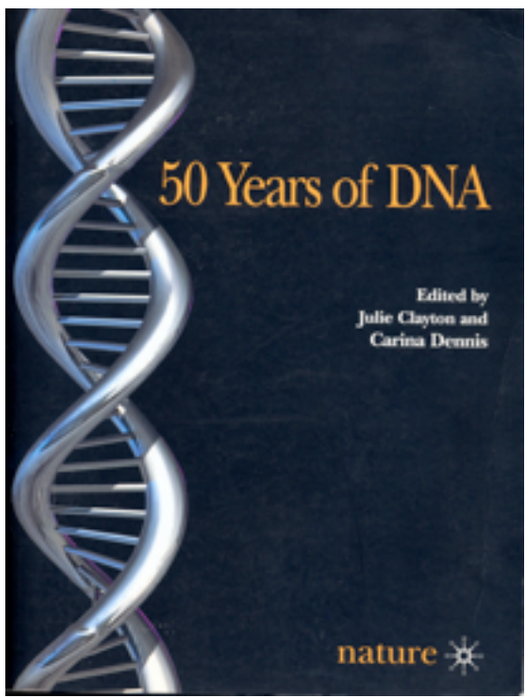
- careful observation
- systematic compilation of evidence
- a logical plan, and
- an understanding that a solution could be obtained from even the tiniest pieces of evidence, extracted from an undisturbed scene of investigation



Natural Digital Preservation

from Wikipedia, Paul Harrison: stromatolites

- DNA as longterm storage
- “Living fossils”
- Conserved DNA



Adapting Existing Technologies for Digitally Archiving Personal Lives

Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools

Jeremy Leighton John

Department of Western Manuscripts, Directorate of Scholarship and Collections, The British Library
96 Euston Road, LONDON NW1 2DB, United Kingdom
jeremy.john@bl.uk

Abstract

The adoption of existing technologies for digital curation, most especially digital capture, is outlined in the context of personal digital archives and the Digital Manuscripts Project at the British Library. Technologies derived from computer forensics, data conversion and classic computing, and evolutionary computing are considered. The practical imperative of moving information to modern and fresh media as soon as possible is highlighted, as is the need to retain the potential for researchers of the future to experience the original look and feel of personal digital objects. The importance of not relying on any single technology is also emphasised.

Introduction

Archives of ‘personal papers’ contain letters, notebooks, diaries, draft essays, family photographs and travel cine films; and in 2000 the British Library adopted the term eMANUSCRIPTS (eMSS) for the digital equivalent of these ‘personal papers’, having begun accepting diverse computer media as part of its manuscript holdings (Summers and John 2001, John 2005).

These media include punched cards, paper tapes, magnetic tapes, program cards, floppy disks of several sizes (8”, 5.25”, 3.5” and 3”), zip disks, optical disks (eg CDRs and DVDRs) and various hard drives, both internal and external. All three major contemporary operating system families are represented: Microsoft Windows, Apple Macintosh and Unix/Linux as well as earlier systems.

Beyond the library’s own collections, the Digital Manuscripts Project has enabled digital capture for the Bodleian Library, the Royal Society (with the National Cataloguing Unit for the Archives of Contemporary Scientists), and the Wellcome Library.

Digital Manuscripts at the British Library

The primary aim of the project is to develop and put into place the means with which to secure the personal archives of individuals in the digital era in order to enable sustained access. This entails the capture of the digital component of the archive alongside its corresponding analogue component.

The project is also addressing in tandem the digitisation of the conventional papers in personal archives (and in that sense is also concerned with digital manuscripts beyond eMSS). Among other benefits, this will make it easier for researchers to work with an entire personal archive in an integrated way; but this work along with cataloguing and resource discovery is beyond the scope of the present paper, which aims to focus on the curatorial role in digital acquisition, examination and metadata extraction.

Theoretical and Practical Considerations

The challenges of technological obsolescence, media degradation and the behaviour of the computer user (eg failure to secure and backup information including passwords) are long familiar to the digital preservation community. Personal collections raise issues, however, that are different from those arising with publications, which have received far more attention.

Of special relevance is the means of acquiring personal archives. Central to the process is the relationship between the curator and the originator or depositor, and in particular the need to deal with personal matters in a sensitive way, ensuring robust confidentiality where necessary.

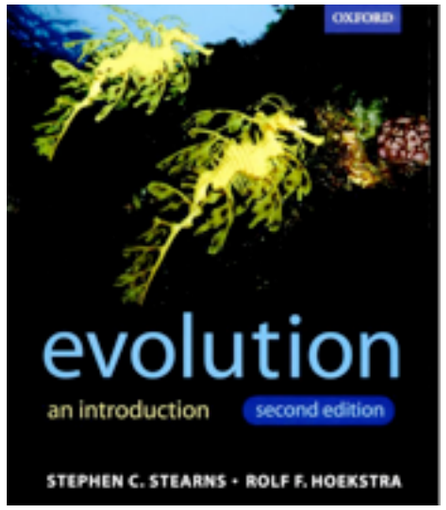
Three key requirements have been identified and promoted: (i) to capture as far as possible the whole contextual space of the personal computer (the entire hard drive or set of hard drives for example) and not just independent individual files, thereby strengthening authentication; (ii) to replicate and retain exact copies of the original files, recognising their historical and informational value (and not just rely on digital facsimiles, even if these match modern standards for interoperability); and (iii) to meet the special requirements for a confidentiality that is sensitive and reassuring to potential depositors as well as being technically convincing.

A pragmatic philosophy is to provide for immediate access to basic text, images and sounds (eg raw alphanumeric content which will suffice for many scholarly purposes); but to retain (by capturing and keeping exact digital replicates of disks and files) the potential to make available high fidelity versions that respect original styles, layout and behaviour.

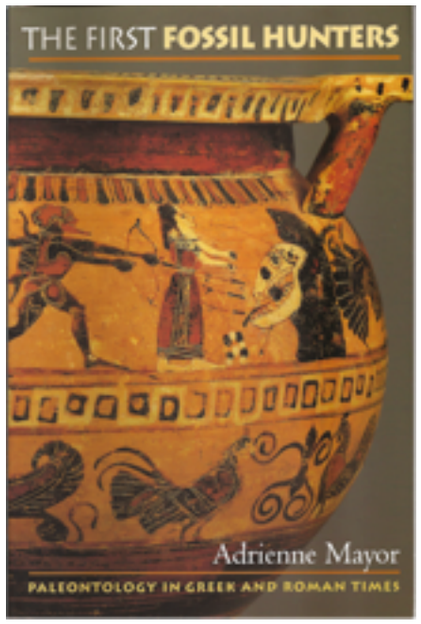
Written in the Rocks

"Paleontologists are the historians of life"

Stephen C. Stearns and Rolf F. Hoekstra (2000) Evolution. An introduction. Oxford University Press, Oxford



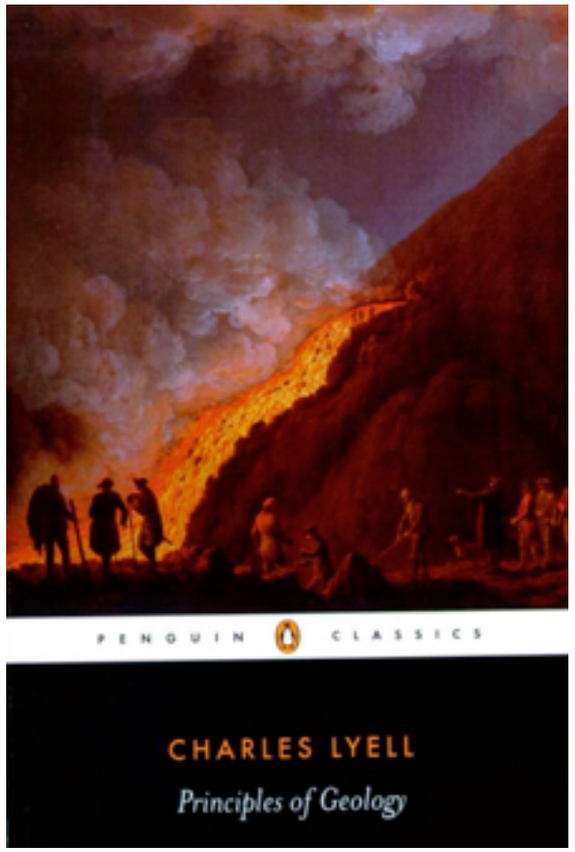
"Just as the fossil is 'petrified time,' so is an ancient artifact or text. The tasks of palaeontologists and classical historians and archaeologists are remarkably similar – to excavate, decipher, and bring to life the tantalizing remnants of a time we will never see."



Adrienne Mayor (2000) The first fossil hunters. Paleontology in Greek and Roman times. Princeton University Press, Princeton

Written in the Rocks

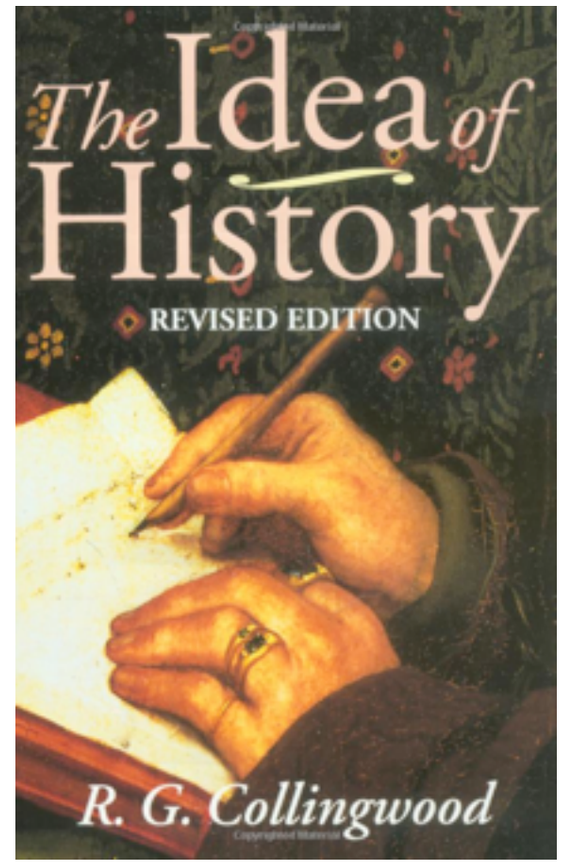
Charles Lyell
Principles of Geology
1830-1833



"When we study history, we obtain a more profound insight into human nature, by instituting a comparison between the present and former states of society. **We trace the long series of events which have gradually led to the actual posture of affairs**; and by connecting effects with their causes, we are enabled to classify and retain in the memory a multitude of complicated relations"

"These topics we regard as constituting the **alphabet and grammar** of geology".

Historical Investigation

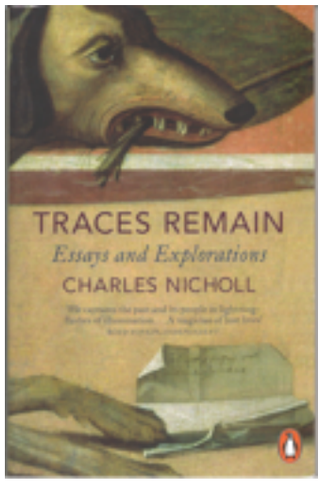


"History, then is a science, but a science of a special kind. It is a science whose business is **to study events not accessible to our observation**, and to study these events inferentially, **arguing to them from something else which is accessible to our observation**, and which the historian calls 'evidence' for the events in which he is interested".

Pp 251-252, The Idea of History by R. G. Collingwood, Revised Edition, 1993, Oxford University Press, Oxford.

Historical Investigation

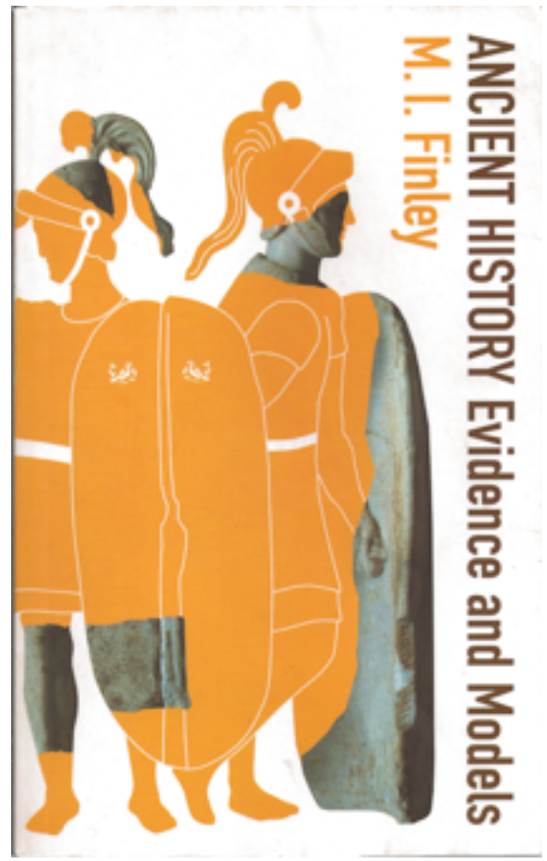
Traces Remain. Essays and Explorations. Charles Nicholl (2011)



“That traces remain is the guiding principle of archaeologists and crime-scene investigators, and though I do not have their particular expertises, I share with them a belief that the evidence of the past, whether distant or recent is something to be examined at close quarters, to be sifted, combed through, picked over”

“... it is the document down in the dusty unglamorous archive which is actually a tangible trace of the past. One reads it as a text, with varying degrees of difficulty, but one also attends to it as a physical relic, the product of a unique conjunction of particularities: a certain person at a certain place and time; their very gesture, and in some cases their very mood and disposition, recorded in those ‘vestiges or marks’ on the paper.”

Historical Investigation



M. I. Finley quotes, p 54, Wilhelm von Humboldt a lecture "On the Historian's Task" given to the Berlin Academy 1821

" 'The historian's task, said Humboldt, 'is to present what actually happened'."

"an event 'is only partially visible in the world of senses; the rest has to be added by intuition, inference, and guesswork… The truth of any event is predicated on the addition… of that invisible part of every fact, and it is that part, therefore, which the historian has to add… Differently from the poet, but in a way similar to him, **he must work the collected fragments into a whole**'."

Natural Retrospection

"Of all natural systems, living matter is the one which, in the face of great transformations, preserves inscribed in its organization the largest amount of its own past history. Using Hegel's expression, we may say that there is no other system that is better *aufgehoben* (**constantly abolished and simultaneously preserved**). We may ask the questions where in the now living systems the greatest amount of their past history has survived and how it can be extracted".

Emile Zuckerkandl and Linus Pauling

Molecules as documents of evolutionary history
Journal of Theoretical Biology 8: 357-366 (1965)

Biohumanities

2009
Volume 3 Number 4

Digital Humanities Quarterly

2009 3.4 | XML | Discuss (0 Comments)

Conjectural Criticism: Computing Past and Future Texts

Kari Kraus

College of Information Studies and Department of English
University of Maryland

Abstract

Broadly conceived, this article re-imagines the role of conjecture in textual scholarship at a time when computers are increasingly pressed into service as tools of reconstruction and forecasting. Examples of conjecture include the recovery of lost readings in classical texts, and the computational modeling of the evolution of a literary work or the descent of a natural language. Conjectural criticism is thus concerned with issues of transmission, transformation, and prediction. It has ancient parallels in divination and modern parallels in the comparative methods of historical linguistics and evolutionary biology.

The article develops a computational model of textuality, one that better supports conjectural reasoning, as a counterweight to the pictorial model of textuality that now predominates in the field of textual scholarship. "Computation" is here broadly understood to mean the manipulation of discrete units of information, which, in the case of language, entails the grammatical processing of strings rather than the mathematical calculation of numbers to create puns, anagrams, word ladders, and other word games. The article thus proposes that a textual scholar endeavoring to recover a prior version of a text, a diviner attempting to decipher an oracle by signs, and a poet exploiting the combinatorial play of language collectively draw on the same library of semiotic operations, which are amenable to algorithmic expression.

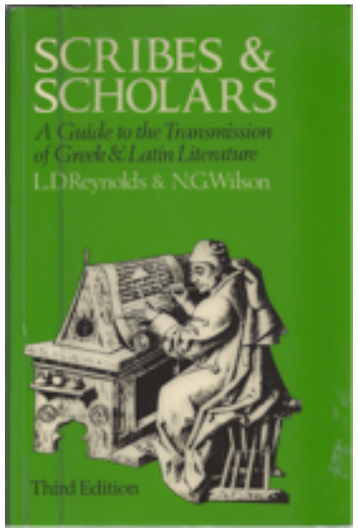
The intended audience for the article includes textual scholars, specialists in the digital humanities and new media, and others interested in the technology of the written word and the emerging field of biohumanities.

“Examples of conjecture include the recovery of lost readings in classical texts, and the computational modeling of the evolution of a literary work or the descent of a natural language”.

The paper notes the early role of textual scholarship in enabling reconstruction, in retrospective analysis

Origins of Stemmatics

L. D. Reynolds & N. G. Wilson, Scribes & Scholars. A Guide to the Transmission of Greek & Latin Literature



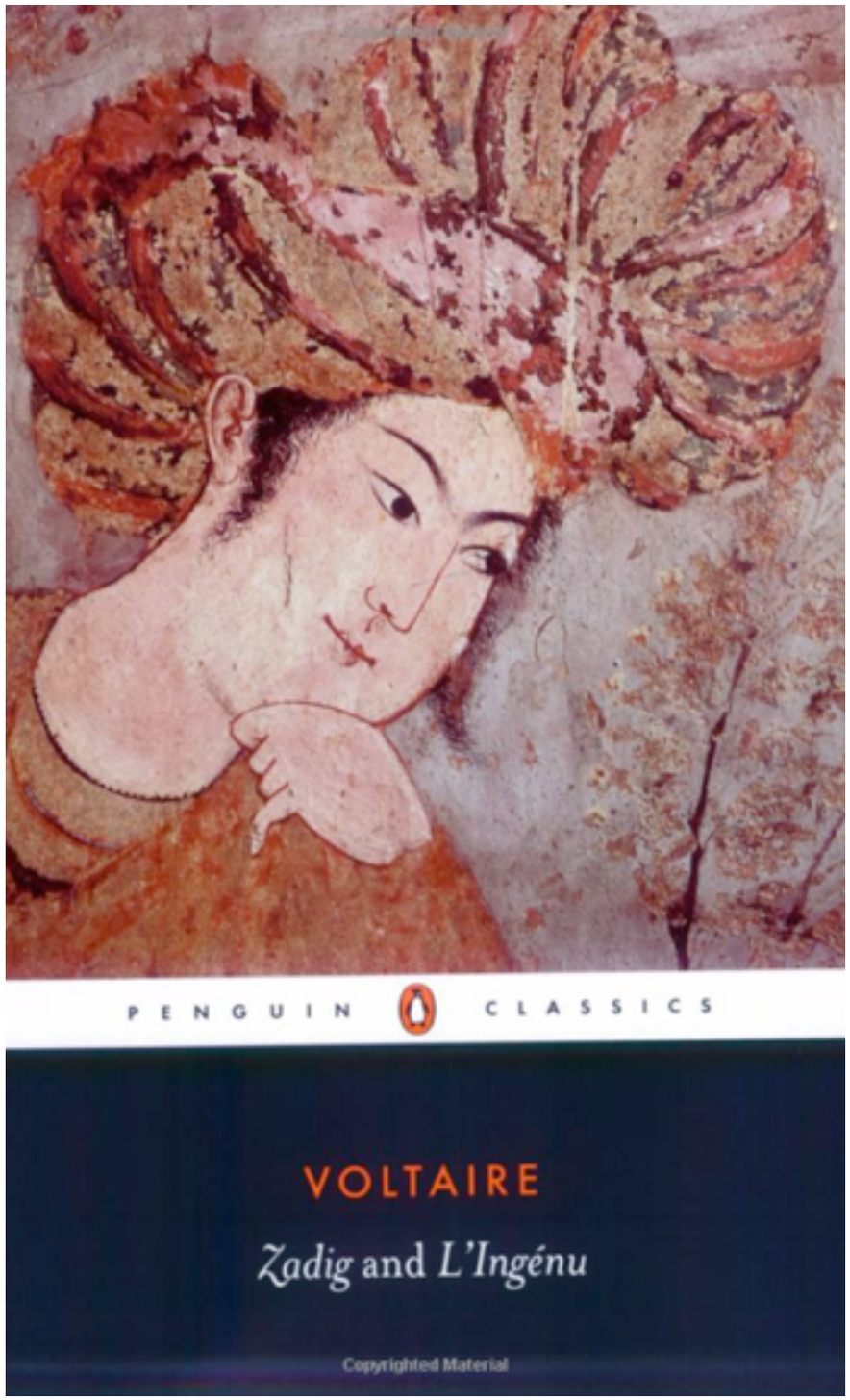
"In 1508 Erasmus postulated a single archetype from which all the surviving manuscripts of a text descended; and although his notion of an archetype was less precisely defined than ours, he was able to explain how easy it is for all the manuscripts to be wrong. The notion of the medieval archetype seems to have been first entertained by Scaliger" [1577].

"Scaliger was far ahead of his time"

BRITISH LIBRARY

Telltale Traces

The Tale of Zadig

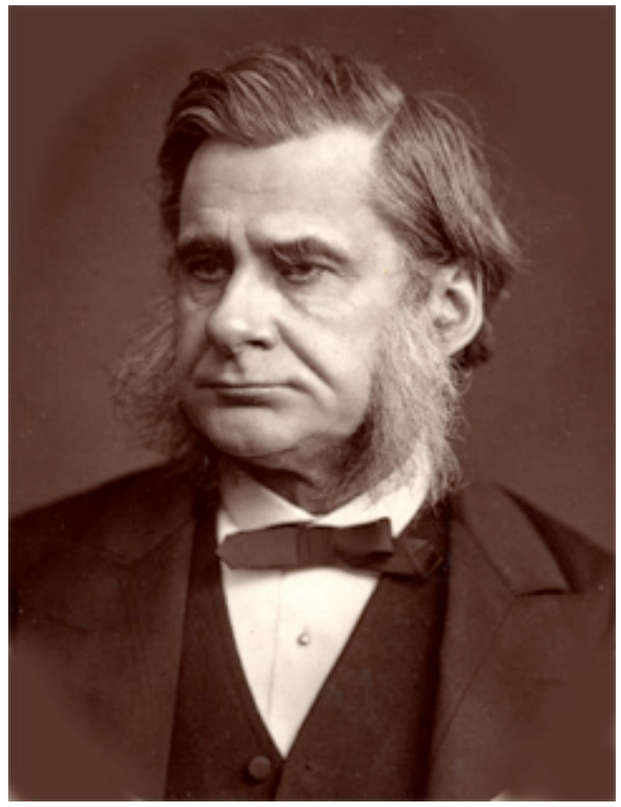


A missing dog and a missing horse

"Other tracks of a different kind, which always appeared to have brushed the sand at either side of the forefeet, showed me that its ears were very long; and as I noticed that the sand was always more deeply impressed by one paw than by the other three, I concluded that our august Queen's little bitch was a trifle lame, if I may dare to say so"

"As for the horse which belongs to the King of Kings, you must know that as I was walking along the paths of this wood, I noticed some horseshoe prints all at equal distance. 'There's a horse with a fine gallop,' I said to myself. **In a straight stretch of path only seven feet wide, the dust had been lightly brushed from the trees on both sides at a distance of three and a half feet from the centre of the path. 'This horse,' said I, 'has a tail three and a half feet long, which must have swept off the dust on both sides as it waved.**' The trees formed an arcade five feet high. When I noticed that some of the leaves were newly fallen, I deduced that the horse must have touched them, and that he was therefore five feet high also. As for the bit, it must be made of twenty-three carat gold, because the horse rubbed the bosses against a stone which I knew to be touchstone, and which I therefore tested. And finally I judged from the marks which the horseshoes had left on a different kind of stone that it was shod with silver of eleven deniers proof"

T. H. Huxley on Palaetiology



"For the rigorous application of Zadig's logic to the results of accurate and long-continued observation has founded all those sciences which have been termed historical or palaetiological, because they are retrospectively prophetic and **strive towards the reconstruction in human imagination of events which have vanished and ceased to be**".

by Thomas H. Huxley (1880) On the method of Zadig. Retrospective prophecy as a function of science. Science and Hebrew Tradition.

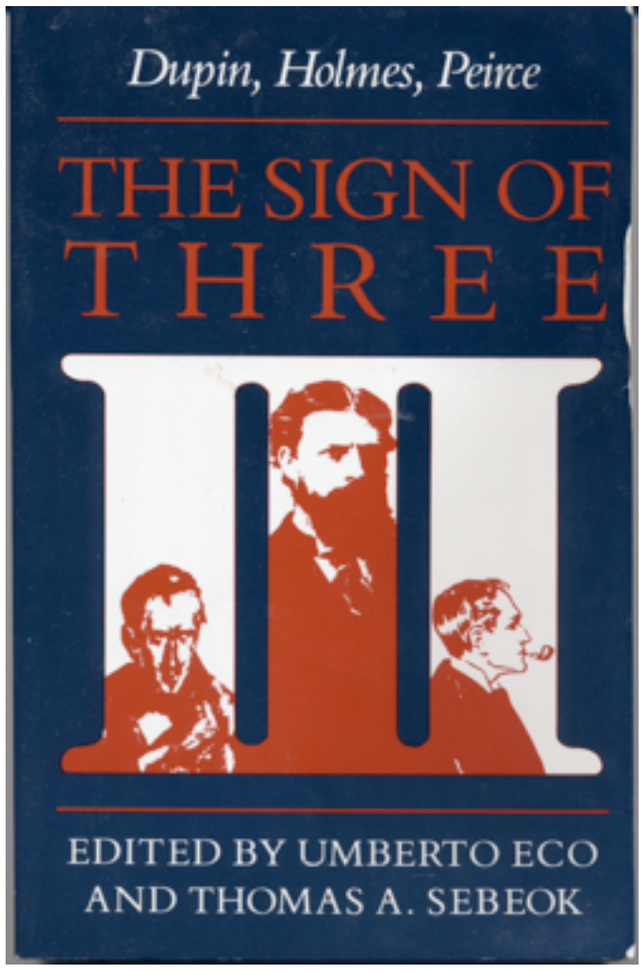
"The tracks were exactly like those which dogs and horses leave: therefore they were the effects of such animals as causes".

"History, in the ordinary acceptation of the word, is based upon the interpretation of documentary evidence; and **documents would have no evidential value unless historians were justified in their assumption that they have come into existence by the operation of causes similar to those of which documents are, in our present experience, effects**"

Images from Wikipedia

Semiotics

Chapter Four, Clues: Morelli, Freud and Sherlock Holmes, pp 81-118, by Carlo Ginzburg

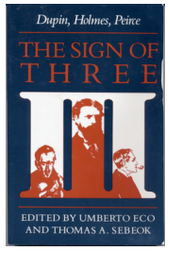


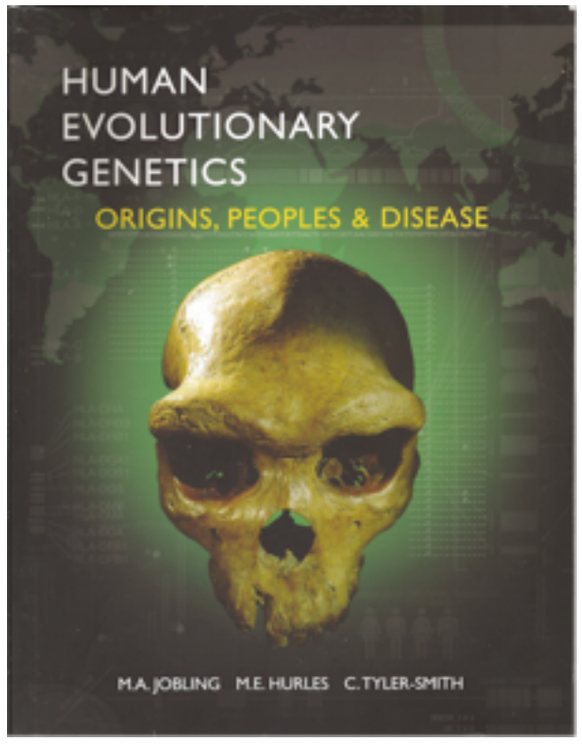
"Perhaps indeed the idea of a narrative, as opposed to spell or exorcism or invocation…, originated in a hunting society, from the experience of interpreting tracks." "The hunter could have been the first 'to tell a story' because only hunters knew how to read a coherent sequence of events from the silent (even imperceptible) signs left by prey".

"This 'deciphering' and 'reading' of animals' tracks is metaphorical. But it is worth trying to understand it literally, as the verbal distillation of a historical process leading, though across a very long time-span, toward the invention of writing. The same connection is suggested in a Chinese tradition explaining the origins of writing, according to which it was invented by a high official who had remarked the footprints of a bird in a sandy riverbank…".

Dupin, Holmes, Peirce. The Sign of Three, edited by Umberto Eco and Thomas A. Sebeok (1983) (Midland Book Edition 1988), Indiana University Press, Bloomington and Indianapolis

Disease: Signs and Symptoms





Chapter Four, Clues: Morelli, Freud and Sherlock Holmes, pp 81-118, by Carlo Ginzburg

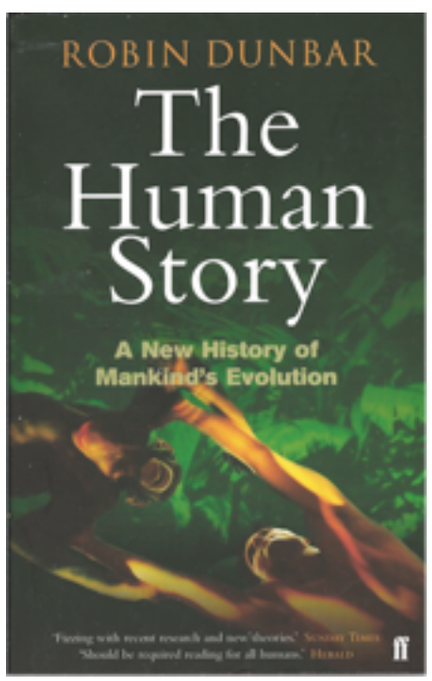
"In all three cases [Morelli, Freud, Holmes] we can invoke the model of medical semiotics or symptomatology – the discipline which permits diagnosis, though the disease cannot be directly observed, on the basis of superficial symptoms or signs, often irrelevant to the eye of the layman…"

Dupin, Holmes, Peirce. The Sign of Three, edited by Umberto Eco and Thomas A. Sebeok (1983) (Midland Book Edition 1988), Indiana University Press, Bloomington and Indianapolis

Mind Reading, Hunting, Toolmaking

A number of intellectually demanding activities, many centred around human understanding of cause and effect, have been raised as possible explanations for the considerable energetic and structural investment in brains in humans and time spent learning:

- toolmaking
- foraging
- hunting
- identification of health and its absence
- social interaction



Recently much of the emphasis has been on social interaction and networking including story telling, ornament creation and the ability to surmise the thoughts and wishes of other people; but activities such as effective toolmaking, hunting and socialisation may not be mutually exclusive

Even forensics is in the final analysis a social phenomenon, a cooperative and adversarial process of thinking

BRITISH LIBRARY

Principles

Principles: Geology

“[James] Hutton... had convinced himself that basalt and many other trap-rocks were of igneous origin, and that many of them had been injected in a melted state through fissures in the older strata”.

Charles Lyell
Principles of Geology
1830-1833

“The first writer to recognize the organic nature of fossil shells was the Greek philosopher Xenophanes”

The fossil record of the Mediterranean lands is extremely fragmented with the disrupted stratigraphy not offering orderly sequences of fossils in layers that would point to a series of animal types in classical antiquity

Adrienne Mayor (2000) The first fossil hunters. Paleontology in Greek and Roman times. Princeton University Press, Princeton

Principles: Geology

Determination of the Relative Age of Rocks

Proofs of Relative Age by Superposition

Proofs by Included Fragments of Older Rocks

Proofs of Contemporaneous Origin derived from Mineral Characters

Proofs of Contemporaneous Origin derived from Organic Characters

Charles Lyell
Principles of Geology
1830-1833

Principles: Stemmatics

"Of fundamental importance in stemmatics are the errors which scribes make in copying manuscripts; for these errors provide the most valid means of working out the relationships of the manuscripts. **Special attention is paid to errors of omission and transposition.**

For stemmatic purposes these errors can be divided into

(a) those which show that **two manuscripts are more closely related to each other** than to a third manuscript (conjunctive errors), and

(b) those which show that **one manuscript is independent of another** because the second contains an error or errors from which the first is free (separative errors).

Care is taken to see that these errors are 'significant', i.e. not such mistakes as two scribes are likely to make independently, or such as a scribe could easily remove by conjecture"

L. D. Reynolds & N. G. Wilson, Scribes & Scholars. A Guide to the Transmission of Greek & Latin Literature

Principles: Forensics

A paradigm of forensic science is offered by Inman and Rudin (2000), centred around the concepts of identification, classification and individualization

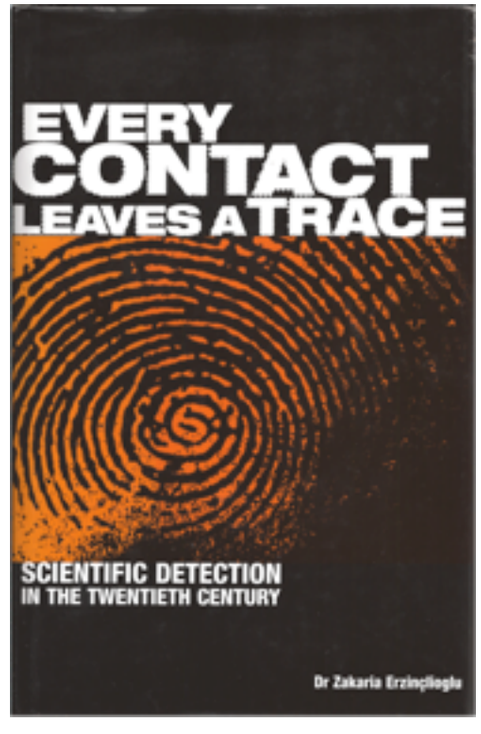
Individualisation refers to the concluding of a single common source for evidence.

It may be distinguished from **classification** where a number of potential common sources may be inferred.

Do these two or more items share a single common origin?

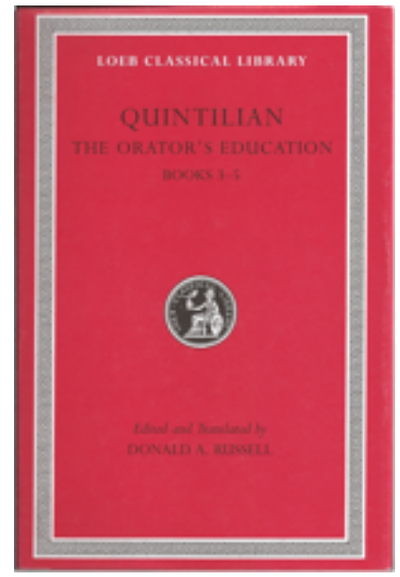
K. Inman and Norah Rudin (2000) Principles and Practice of Criminalistics. The Profession of Forensic Science

Principles: Forensics



The most well known principle is attributed to Edmond Locard (1877–1966) the Exchange or Transfer Principle that ‘every contact leaves a trace’

The Multievidential Perspective



"But it may be the blood of a sacrificed animal that has got onto the clothes, or just a nosebleed: a man whose clothes are bloody has not necessarily committed homicide. **But though this Sign is not enough in itself, in combination with others it is taken as a piece of evidence**, if the man is an enemy, or has previously made threats, or was in the same place"

"The force of these Signs depends entirely on the support they have from other sources."

Quintilian. The Orator's Education, Edited and translated by Donald A. Russell (2001) Books 3-5, Book 5.9

From the modern perspective, this requirement for multiple forms of independent data is a central aspect of statistical theory and the way Bayesian analysis, for example, builds up evidence

BRITISH LIBRARY HISTORIES

Phylomemetics

Phylogenetics, Stemmatics, Detection

DEEP PAST

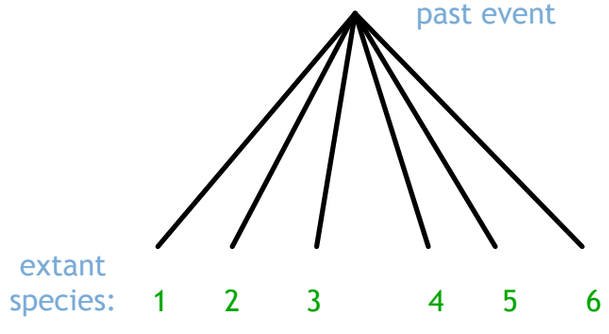
DISTANT PAST

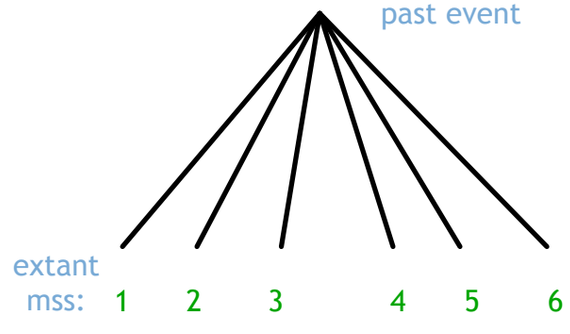
RECENT PAST

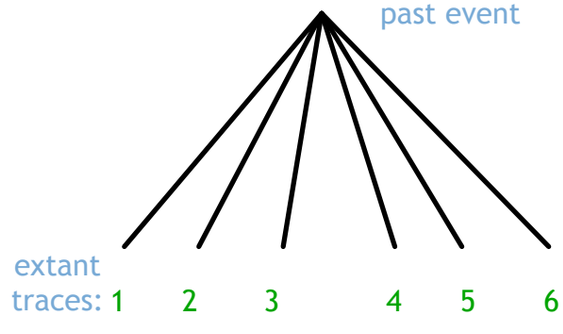
evolutionary biology

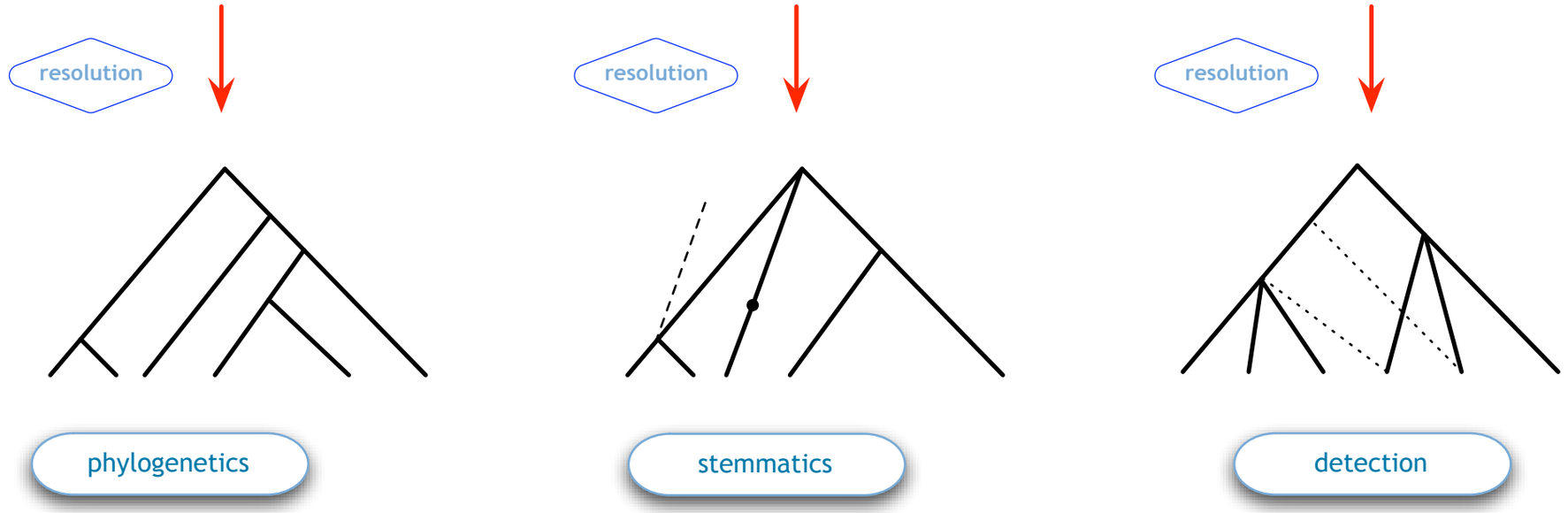
historical research

forensic science







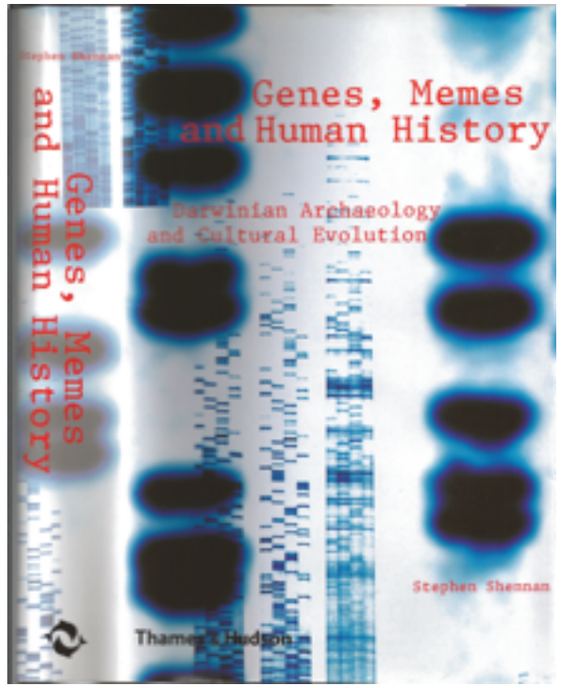


graphic: digital lives: jeremy leighton john

Approaches towards reconstructing the past

Cultural Diversity and Human History and Prehistory

S. Shennan (2002)

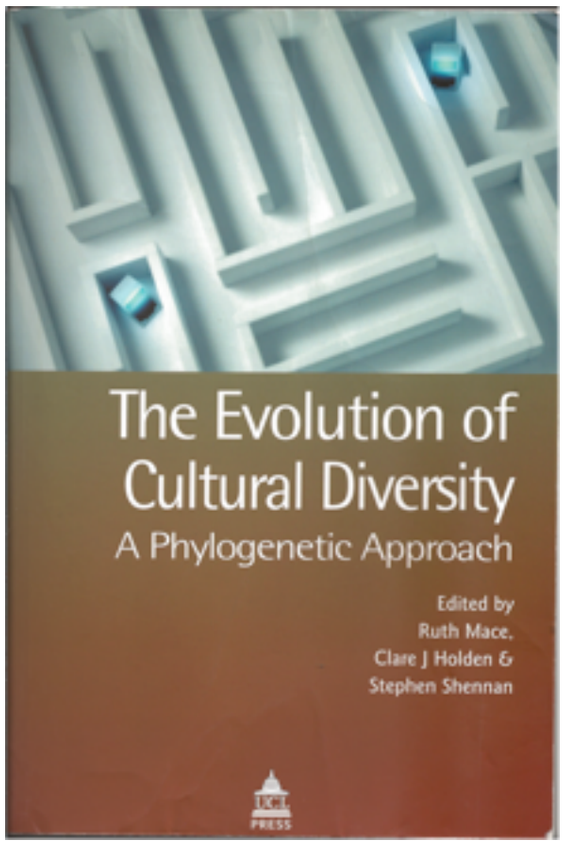


The orientalist William Jones (1746-1794) proposed the idea that Sanskrit and Greek were derived from a common ancestor, and that there is a familial or genetic (in the early sense) relationship between certain Indo-European languages involving both inheritance and change.

Palaeontologist R Foley (2006) concluded: “Stone tools make up the richest part of our evolutionary past and extend the human fossil record beyond the morphological to the behavioural”.

Phylogenetic approaches are being adopted by social and physical anthropologists as well as archaeologists and are being applied across a range of fields from **language, textile patterns, musical instruments, canoe design, games, basket structure, and Mycenaean Linear B script**

R. Mace, C. G. Holden and S. Shennan (2005)

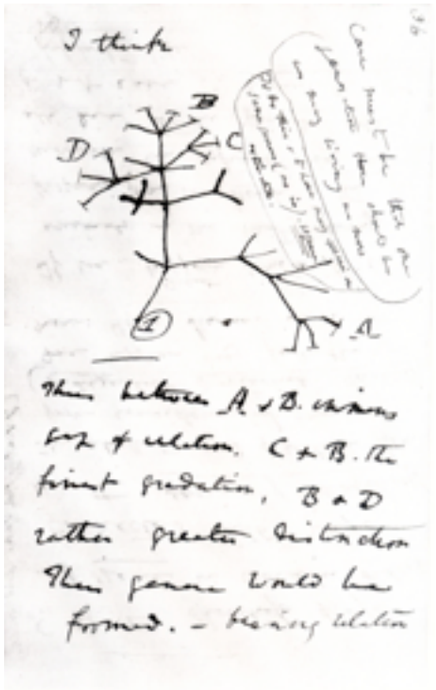


Artefacts and Language

Darwin and many of his contemporaries believed that the analogy between organisms and languages could be extended to other cultural domains

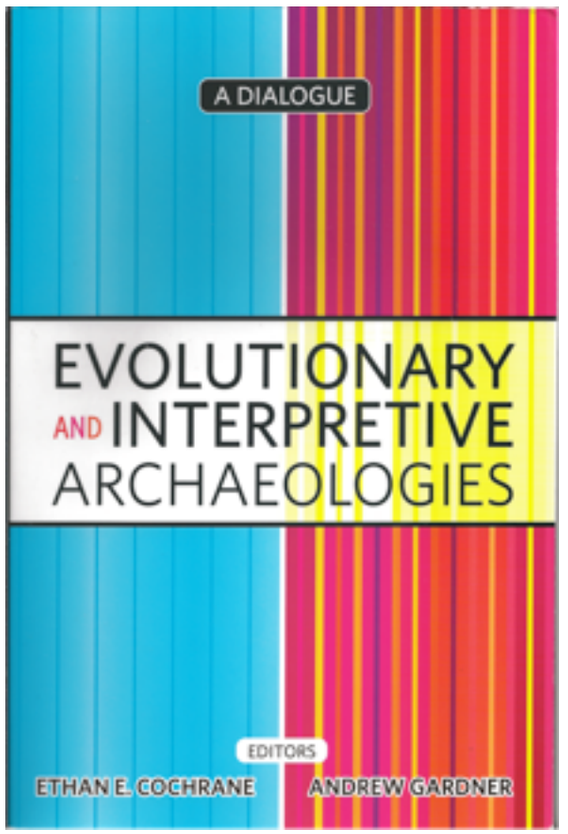
One of the most important of these figures was Augustus Henry Pitt-Rivers (Pitt-Rivers 1875, 1906), whose ethnographic collections were conceived with the express intention of demonstrating how the principles of evolution are borne out in tools, weapons and craft objects”

“Thus, he describes how artefacts are generally copied from other artefacts, just as biological individuals are copies of other individuals. He also noted that copies may differ slightly from their model due to small errors or improvements in design and technique…”.



Charles Darwin 1837 Notebook

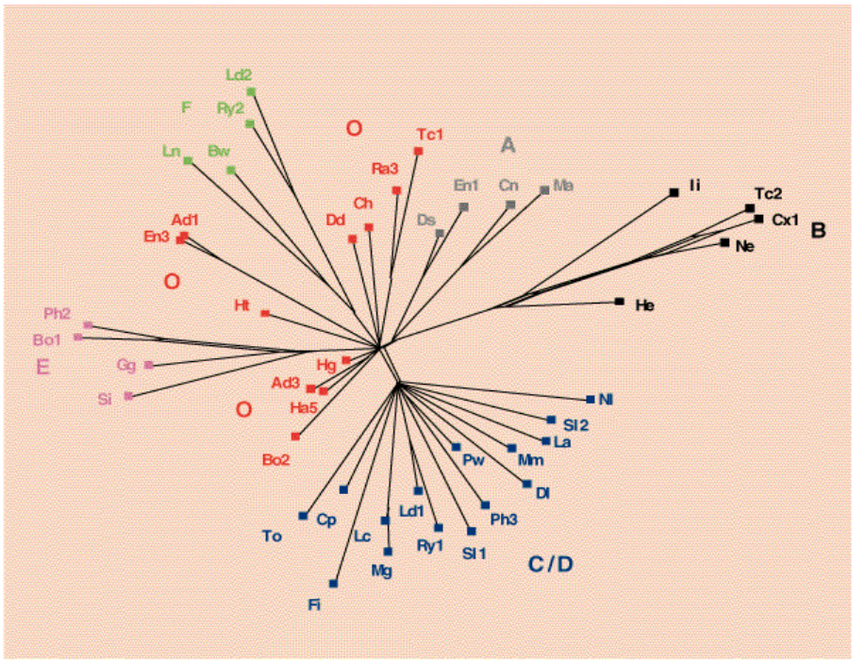
Charles Darwin: language
August Schleicher: language Indo-European



Jamshid J. Tehrani pp 245-261 Missing links: cultures, species and the cladistics reconstruction of prehistory

Chapter 11, in Evolutionary and interpretative archaeologies. A dialogue. Edited by Ethan E. Cochrane and Andrew Gardner (2011), Left Coast Press, Walnut Creek, California

Phylogenetic Trees & Literature



from Barbrook, Howe, Blake & Robinson 1998
Nature 394: 839-840

Textual scholars reconstruct historical relationships between variants that have arisen through copying and modifications made by scribes (deliberately or erroneously). A phylogenetic analysis of 58 extant fifteenth century manuscripts of ‘The Wife of Bath’s Prologue’ from The Canterbury Tales, yielded a phylogeny showing ancestral relationships among the manuscripts with some being closer to the ancestor (A C Barbrook et al., 1998). To strengthen the approach tests have been conducted with artificial manuscripts (eg M Spencer et al., 2004).

Phylogenetic Principles: Methods

“Cladistic analysis reconstructs relationships among taxa or classes by distinguishing between characters that are evolutionarily novel (also termed ‘apomorphic’ or ‘derived’), from those that were present in the last common ancestor of all the taxa under study, which are labelled ‘ancestral’ or ‘plesiomorphic’.”

Several methods including Outgroup Analysis

“Since the outgroup does not share an exclusive common ancestor with any individual member of the ingroup, it follows that when a character occurs in two states among the study group, but only one of the states is found in the outgroup taxon, the former is considered the derived state and the latter the ancestral state.”

Parsimony Principle

The cladistics approach generates a “consensus cladogram that is consistent with the largest number of characters and therefore requires the smallest number of evolutionary changes to account for the distribution of character states among the taxa”

Jamshid J. Tehrani pp 245-261 Missing links: cultures, species and the cladistics reconstruction of prehistory

Phylogenetic Principles: Intentionality

There is a process of continuity, variation and selection at work

A notable distinction lies in the nature of the variation: **the difference between 'erroneous' chance variation and directed variation**. Thus variation of computer viruses, for example, is largely due to the direct work of people modifying existing ones.

Nonetheless there is chance variation as evidenced in historical manuscript copying that is error prone (C J Howe et al., 2001).

A number of researchers have suggested that what matters to the selection process is that variation is available for selection to take place

Phylogenetic Principles: Horizontal Transmission

"The apparent simplicity and finality of the stemmatic method as outlined... is deceptive". "The theory assumes that readings and errors are transmitted 'vertically' from one manuscript to another, that is to say directly from one book to the copies that are made from it."

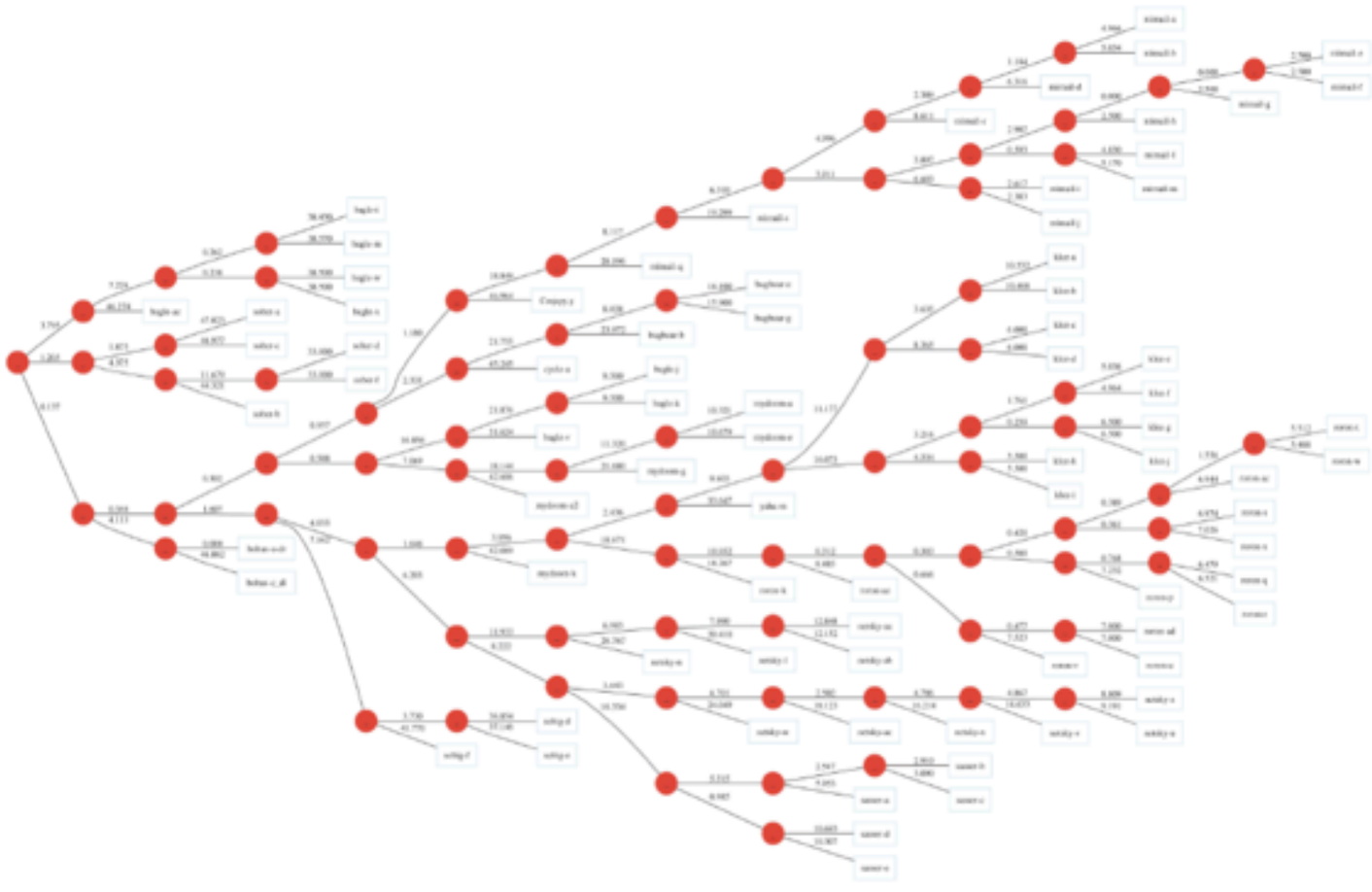
"Readers in ancient and medieval times did not necessarily copy a text from a single exemplar".

Scribes & Scholars

"Tree algorithms attempt to recover evolutionary trees from the data, and will generally work well where languages have split from each other and subsequently have developed along their own paths. If the data are by nature not treelike (e.g. because languages have exchanged traits), then a tree applied to the raw data is not appropriate and either a judicious data coding is required… or a network algorithm is indicated…".

Colin Renfrew and Peter Forster (2006)

Phylogeny of Computing: Malware



from Carrera & Erdelyi 2004 Virus Bulletin Conference September 2004: 187-197

Phylogenetic approaches have been adopted in digital forensics in seeking to understand the origins and taxonomic identity of computer viruses, plagiaristic copying, and other tasks (L A Goldberg et al., 1998, E Carrera and G Erdelyi, 2004)

Phylogeny of Computing: Images

Drafts and the Creative Process

Notions of transformation and transmission can be applied to digital images, both still and moving

Although a photograph often begins with the capture of a single passing irretrievable moment, the photographer commonly subjects that single image to editing and manipulation, yielding a trail of variants.

Similarly, video filming almost always involves extensive editing of such significance that post production is a major phase of the creation of a film

Phylogeny of Computing: Images

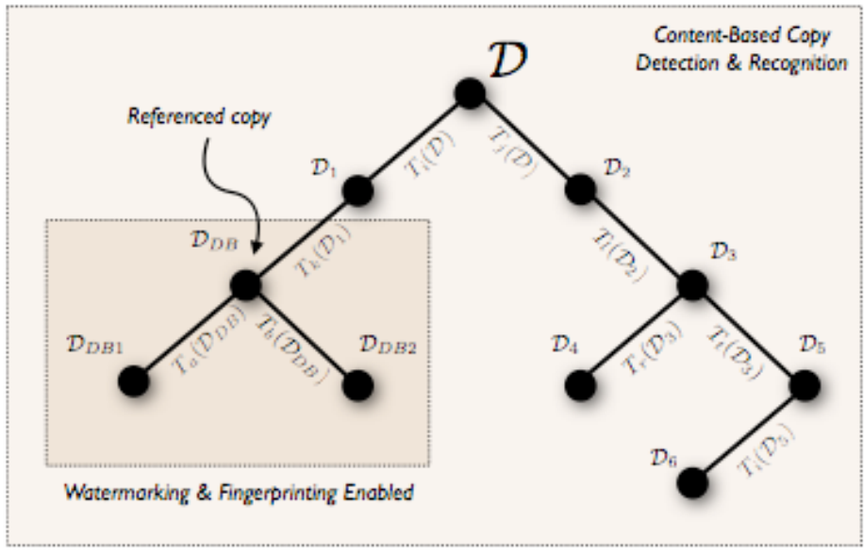


Fig. 1. Near duplicates tree of a document $\mathcal{D}$ and its transformations according to Joly et al [4]. If we can embed a marking on a given document, we can track its transformations easily. On the other hand, when no markings are available or possible, we can use content-based copy detection & recognition methods.

Dias, Rocha and Goldenstein want to establish the history of the duplications of digital photographic objects, digital images, ascertaining “which document generated the other and so on without making use of any watermarking or fingerprinting method”.

A specifically difficult task is the identification of the original digital object, and it may be necessary to undertake this process “even when some pieces, or connections, are missing”

Zanoni Dias, Anderson Rocha and Siome Goldenstein (2010) First steps towards image phylogeny. WIFS 2010, 12-15 December 2010, Seattle, WA

Zanoni Dias, Anderson Rocha and Siome Goldenstein (2011) Image phylogeny by minimal spanning trees, IEEE Transactions on Information Forensics and Security, August 2011

Phylogeny of Computing: Images

"Given a set of *n* near-duplicate images, our first task toward building the image phylogeny tree is to calculate the dissimilarity between every pair of such images. For that we need to consider a good set of possible image transformations, *T*, from which one image can generate a descendant."

Image transformations considered:

Resampling
Cropping
Affine Warping
Brightness and Contrast
Lossy Compression

Zanoni Dias, Anderson Rocha and Siome Goldenstein (2010) First steps towards image phylogeny. WIFS 2010, 12-15 December 2010, Seattle, WA.

Segmentation

TextTiling, an algorithm devised by M. A. Hearst, originated with the dividing of textual documents and the identification of subtopic shift. Roughly, the TextTiling algorithm works by decomposing the entire document into blocks and establishing similarity measures for adjacent blocks. Segment boundaries are supposed where adjacent blocks are of minimal lexical similarity (eg words in common).

Segmentation has been applied to a series of events using a version of the TextTiling algorithm; thus a lifelog of information from a SenseCam and other devices (digital camera, audio recorder, accelerometer, thermometer, infra red and light sensors) may be treated as a ‘document of daily activities’ to be divided into manageable segments, or meaningful events

Aiden R. Doherty and colleagues

Phylogeny of Computing: Software (and Hardware)

Code Conservatism, Code Erosion: Computer Evolution

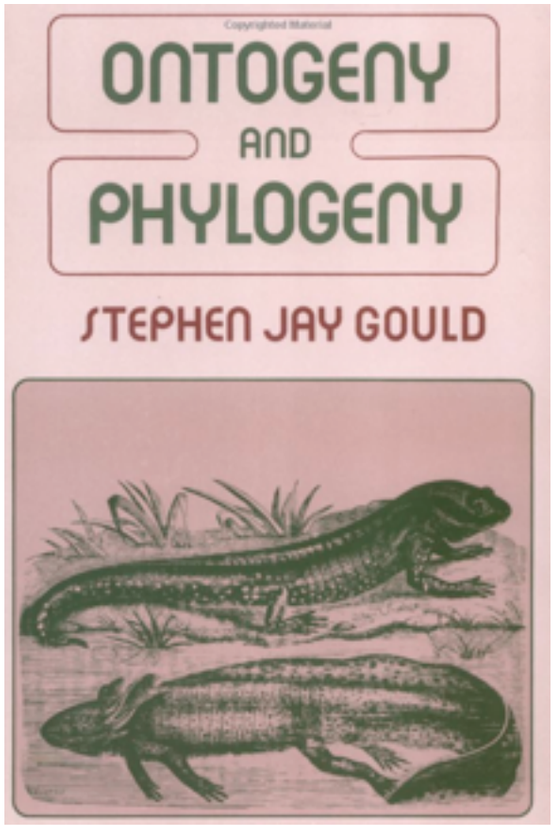
In depth phylogenetic analysis of computer hardware, firmware and embedded code would undoubtedly be revealing. It might, for example, shed light on the way physical features and software code have been shared and borrowed across computer hardware, operating systems and applications.

A number of programming languages share similarities originating in earlier languages.

Some work with software evolution has already proved to be useful in a somewhat different context: code erosion.

C. W. Fraser (2005), “Difftree: Inferring phylogenies for evolving software” Microsoft Research, Technical Report

Computing: Ontogeny Recapitulates Phylogeny?



Legacy Code: Manifestation in Booting Up

Image from Wikipedia

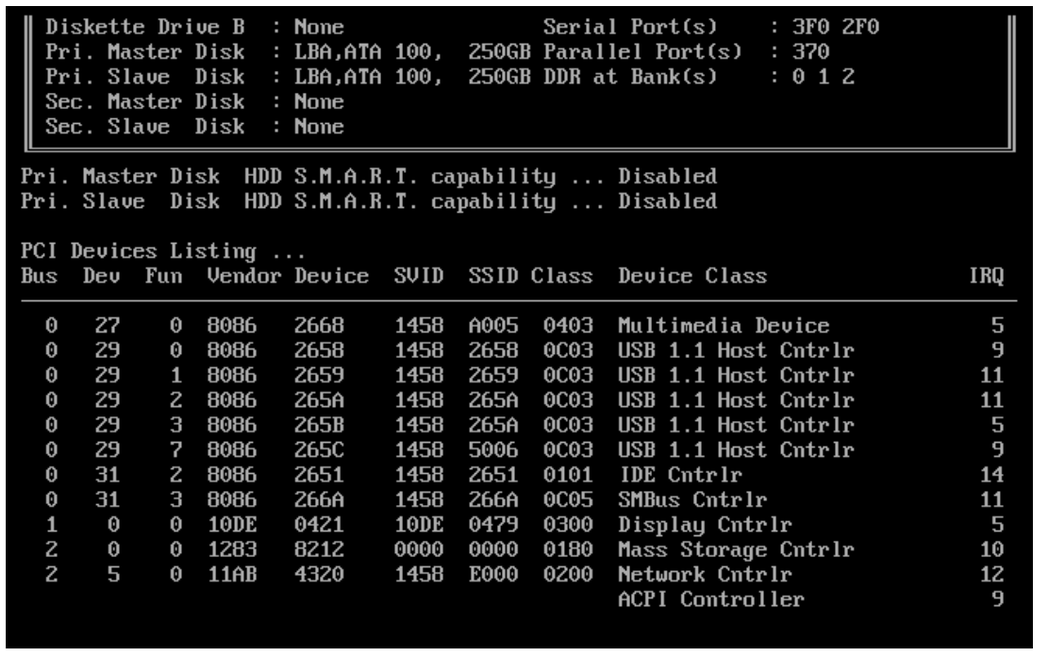


Image from Wikipedia

In order to be backward compatible, some processors still mimic the memory design of the original 8086 family, and when first turned on the computer memory exists in the original ‘real-mode’ restricted to 1 MB (and hence the very basic text presentation)

Personal Digital Archives in the Wild

Personal Digital Archives and Objects

ESSAY

The future of saving our past

As letters and diaries give way to e-mails and laptops, fresh challenges and opportunities have emerged for archivists. **Jeremy Leighton John** explores the digital wilderness for the British Library.

In 2000, the tenaciously original evolutionary biologist and Royal Society professor Bill Hamilton died after an expedition to the rainforests of Africa, and within months his archive was delivered to the British Library in London. Unlike archives received in the past, this one did not consist only of boxes of papers — handwritten letters, typed draft essays and the like. It also included a hoard of computers and storage media from the early 1960s onwards — from 5-hole paper tapes and 80-column punched cards through to optical disks — which occupied 26 of the 200 boxes.

Bill was my mentor, and I grew to know him well, living with his family for four years while completing my PhD in evolutionary biology. At the library, his archive came to me, along with the question of how best to deal with its digital elements. Over the years, Bill's digital archive has been joined by those of other eminent researchers, including evolutionary biologist John Maynard Smith, developmental biologist Anne McLaren and computer scientist Donald Michie.

Digital revolution

The digital revolution is transforming the nature of personal archiving: from curation techniques to the kinds of lives being preserved for posterity — not just the rich and famous, but now everyone participating in the digital age. Most surviving ancient and medieval documents relate to legal, ecclesiastical and formal secular knowledge. Few bear direct witness to the details of everyday lives. Since the sixteenth and seventeenth centuries, paper has helped to open the way for widespread personal writing, printing has facilitated greater literacy and a more robust postal system has promoted distant communication. Even so, unless an archive has gone to a repository or collector, personal papers have been kept in numbers by only the wealthiest families.

With the emergence of personal computing in the 1970s, more and more people are passing on details of their lives to future generations as digital files. It has been estimated by the International Data Corporation that by 2010

ROSEMARY WOODS

nearly 70% of the digital universe will be created by individuals rather than organizations. This offers an immense source of fascinating and useful personal data. But it also presents technical, ethical and legal challenges.

In 2003, I was formally appointed as the first curator of what we call eMANUSCRIPTS at the British Library, responsible for these personal digital objects and for the digital manuscripts project, which aims to research and develop the necessary procedures to curate such archives.

How can the authenticity of files be demonstrated years and centuries later? How can we ensure that a file's dates and other embedded metadata are not inadvertently changed, and are properly extracted? These kinds of considerations motivated the use of computer forensic techniques and the establishment of a secure digital scriptorium in the British Library. The facility is equipped with forensically sound computers and devices that can make exact copies of disks — an approach that is now being adopted by archival institutions worldwide.

The original files residing on a scientist's disks and tapes will eventually become unreadable. It is therefore imperative to make exact copies of these files on fresh media in a demonstrably rigorous way. Merely turning a computer on risks altering files, so devices called write-blockers are used to prevent this from happening.

At the heart of the process is the forensic 'imaging' of an original disk. A single 'image' file, or bitstream, is made of the entire disk. It incorporates a 'map' that allows exact digital replicates of the original files to be recreated from it. Hash values are calculated for every file: short sequences of alpha-numeric characters akin to 'digital fingerprints', unique to each file. If, decades later, the same hash value is obtained for a file using the same hashing algorithm, one can be confident that the file has not been altered.

So as not to rely on software that is subject to obsolescence or is esoteric, we also create (with varying degrees of fidelity to the look and feel of the original) digital facsimiles that, through conversion to suitable file formats such as PDF or XML, are readily usable on modern computers and (let us hope) easily transferred to new computer systems.

However, the digital replicates must be preserved to retain — to the fullest extent possible — the information represented in the original files. This information will be needed should, for example, a future scholar wish to interpret precisely and accurately the original styles, layout and dynamic behaviour of a scientist's files — including home-made computer programs. These replicates can be presented with high-fidelity emulators of ancestral software and hardware.

The British Library houses a small range of classic computers with tape and disk

DIGITAL >>LIVES

An Initial Synthesis

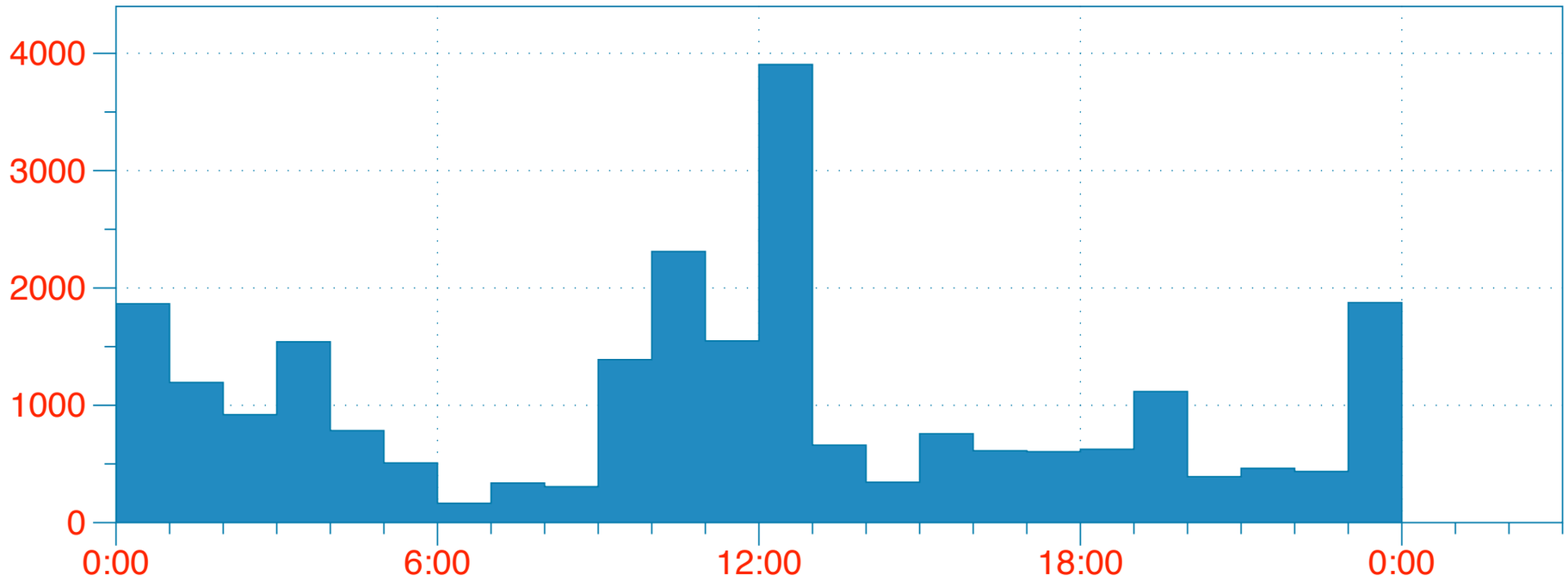
Personal digital devices

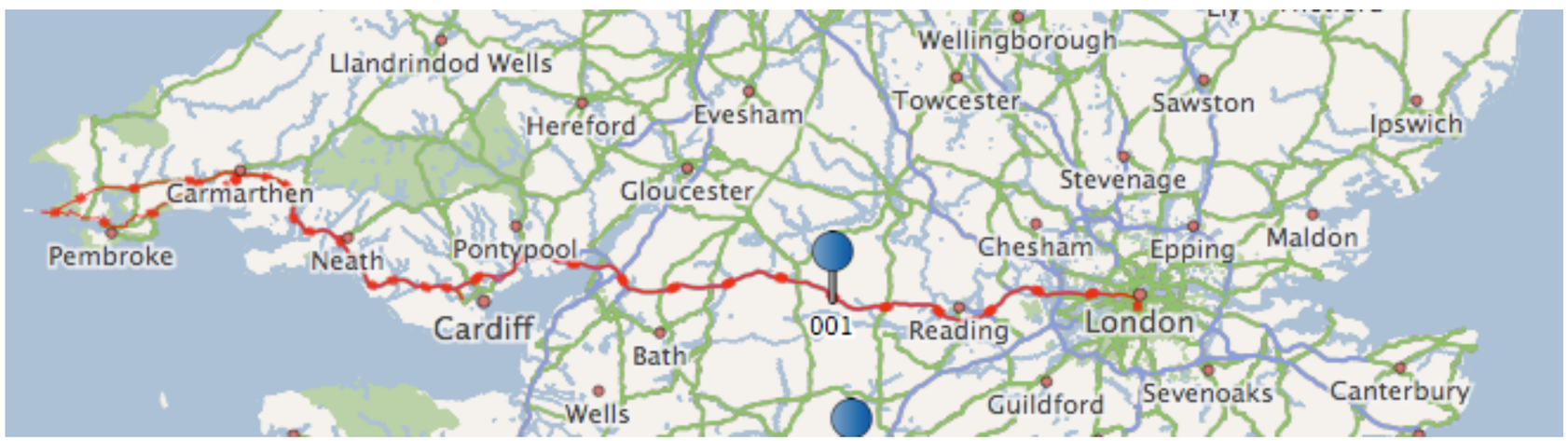
The personal digital archive helps enable

USABILITY

which in turn motivates the personal digital archive?

Personal Informatics: Time and Place





eMSS, Memes, Genes

Personal Digital Archives and Objects

With a paper letter or diary only one sibling could inherit the original from a parent, but today siblings are receiving identical personal digital objects (from texts to videos).

After many generations, one might expect highly diverse personal digital archives distributed throughout the population, containing many replicates of identical personal digital objects (as well as versions that have been modified, deliberately or inadvertently).

It will may be possible to create phylogenetic networks or trees from these extant personal digital archives, and even to surmise from them the composition of ancestral archives.

In the digitally networked era it may be possible for the first time to capture, map and analyse in powerfully quantitative detail the movement and influences of personal digital objects with the ideas, observations of nature and accounts of events borne by them

Phenotypic Expression

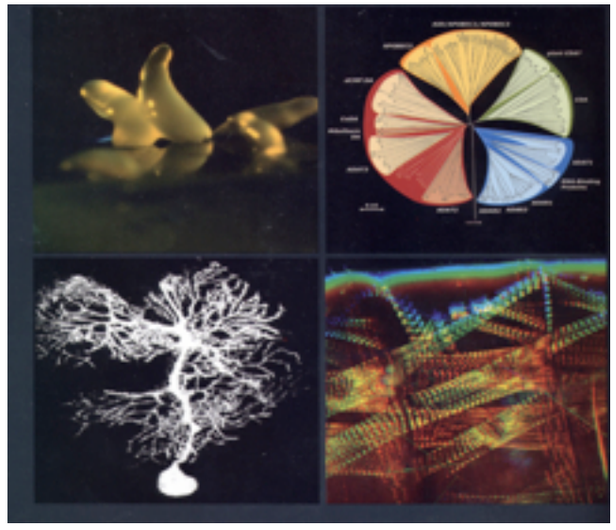
Phenotypic expression of the eMSS might be seen simply as rendering of the digital object or as the processes which they instigate, notably the use made of the eMSS, which is in a sense a central impact of the information represented by the eMSS in the environment

There is a sense in which eMSS may not only instruct but also may impel, incite or instigate, semantically or otherwise

With the arrival of 3D printers, it is possible to envisage that individuals will design personalised physical artefacts (from furniture to ornaments) and 'print' them through a remote service that delivers the artefacts to the doorstep

Also relevant may be the emerging possibilities of:

networks of everyday devices (the so-called 'internet of things') and of ubiquitious computing with devices (including haptic actuators) tuned to the personal experiences and preferences of individuals



Natural Systems Biology & Ubiquitous Computing

"In general, we need a way to model, engineer and understand a society of devices embedded in a society of humans, embedded within a real physical world. Such models would go far beyond the current distributed engineering frameworks, viewpoints and transparencies. Once such models are uncovered, they may be mapped into design patterns – many of these may resemble natural systems. Hence, interdisciplinary research with natural biology may be an essential element for this challenge"

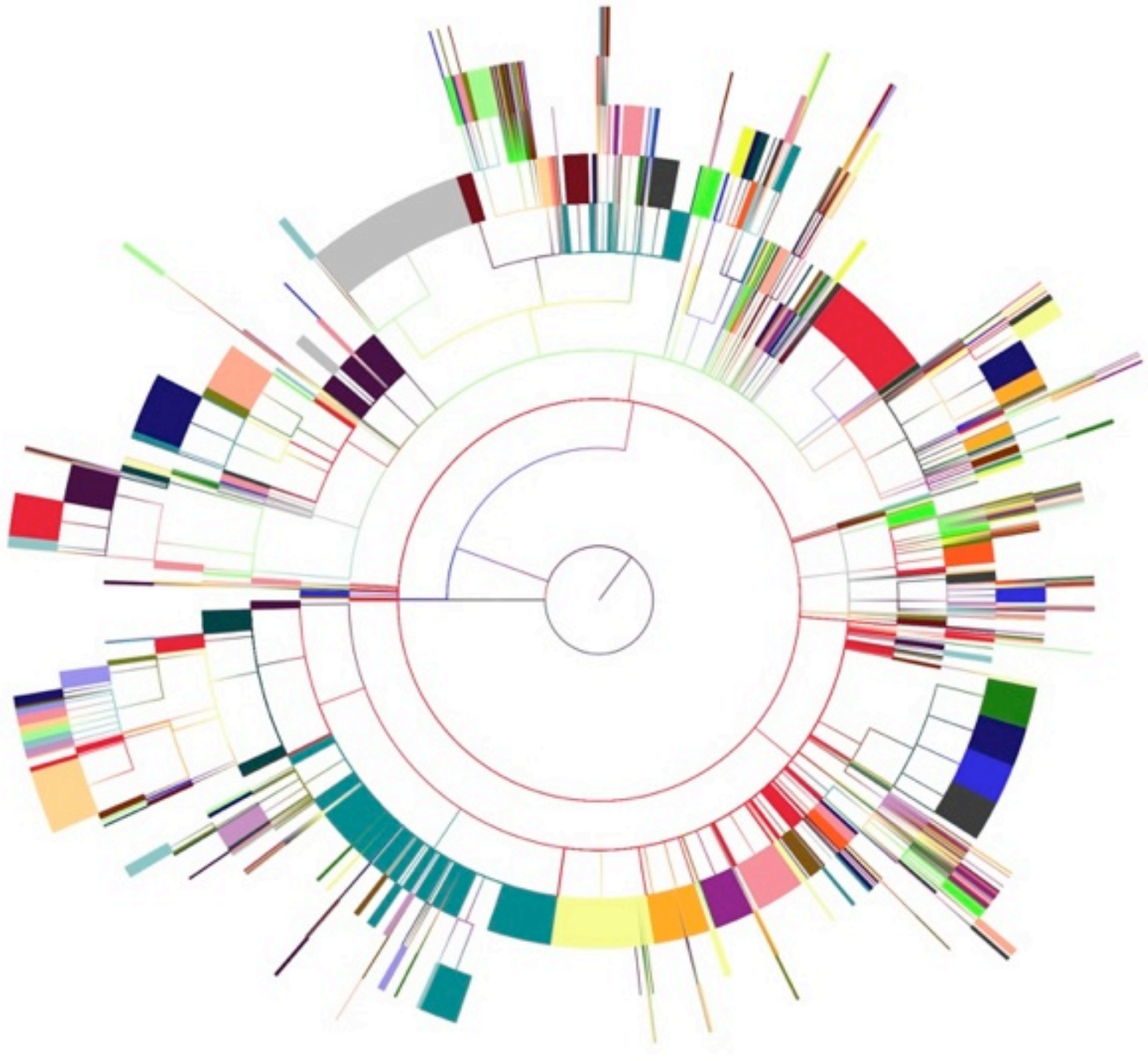
J Crowcroft, 2008

BRITISH LIBRARY

Tools

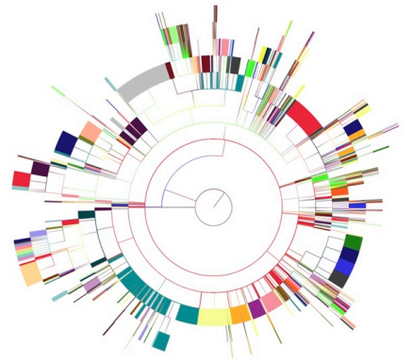
BRITISH LIBRARY

Tree Visualisation



File and folder tree of nearly nearly 14,000 files from one of the hard drives of John Maynard Smith at the British Library using FigTree

Tree Visualisation and Annotation



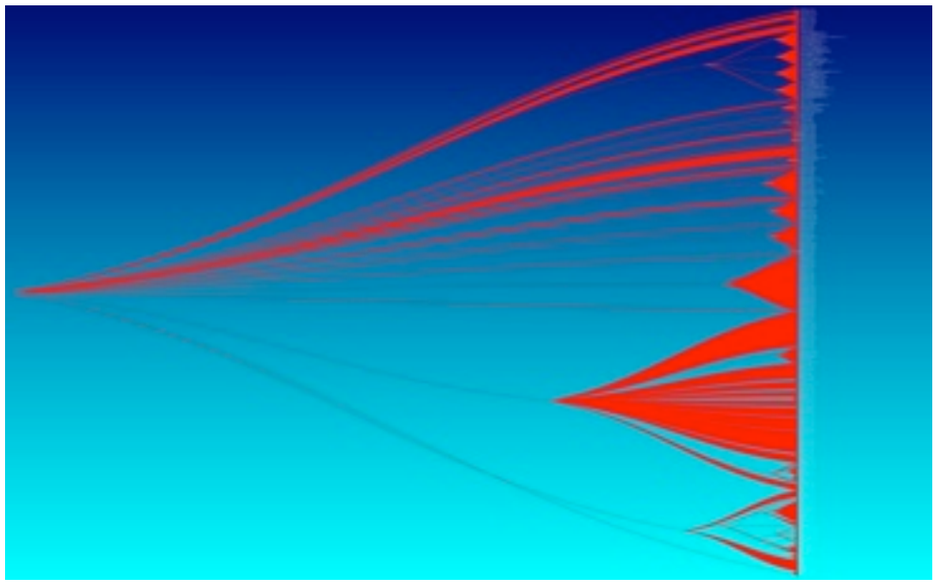
Forensic techniques make it possible to extract large volumes of information from the storage media of personal computers. A subsequent challenge lies in the effective presentation of this information. Paper archives are commonly represented by a set of arrangement records, which taken together essentially map the way the papers such as letters and notes were originally held in envelopes within folders within bundles of folders. Computer files are similarly arranged in a logically hierarchical system within digital media.

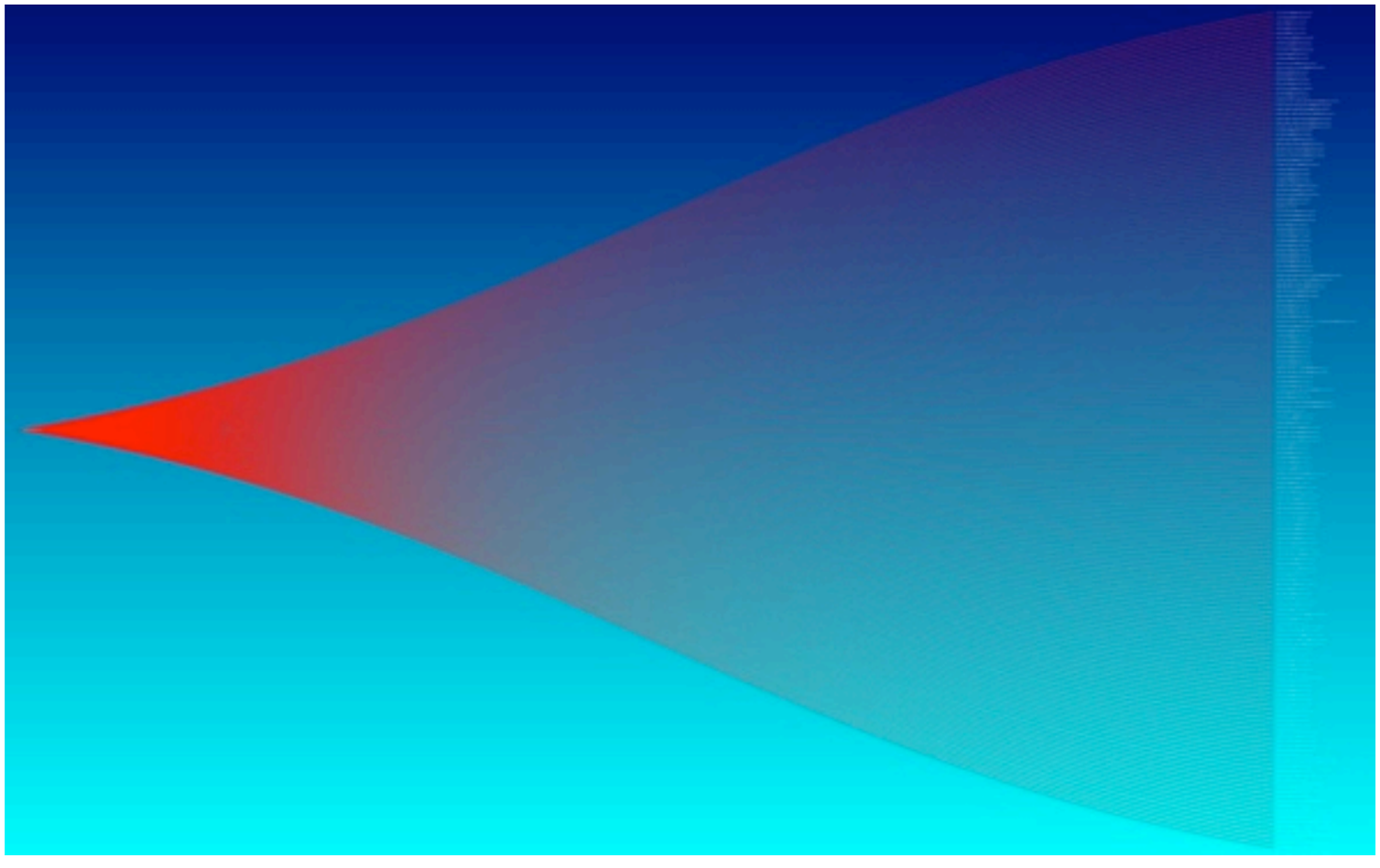
Phylogenetic software has been designed to show the evolutionary relationships between living organisms, and can enable annotation at each internal node (the nodes akin to folders) as well as at the leaf nodes (the species or entities at the tips of the branches).

Annotation may take the form of core metadata created automatically by a program, and supplementary metadata compiled by a scientist or curator, and there may be links to pertinent ancillary information or objects including digital images of the species.

Tree Visualisation

The software Archaeopteryx is able to handle large trees and may be used to convert Newick into other forms such as phyloXML which can in turn be viewed and edited using an XML editor





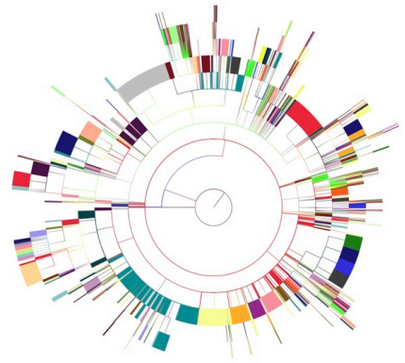
Two low resolution screenshots of the file tree of the same hard drive using Archaeopteryx. The right one is the result of clicking on one of the nodes in the left one. This is one of several ways of quickly navigating the tree

TreeCurator

Usually phylogeneticists obtain their Newick file directly from the software that undertakes the phylogenetic analysis. In order to use phylogenetic tree viewers in another context it is necessary to create the Newick file independently.

The eMSS Lab at the British Library has been writing programs in Python for creating the necessary files in Newick format, and they may be seen as an initial component of a tool to be known as TreeCurator

TreeCurator

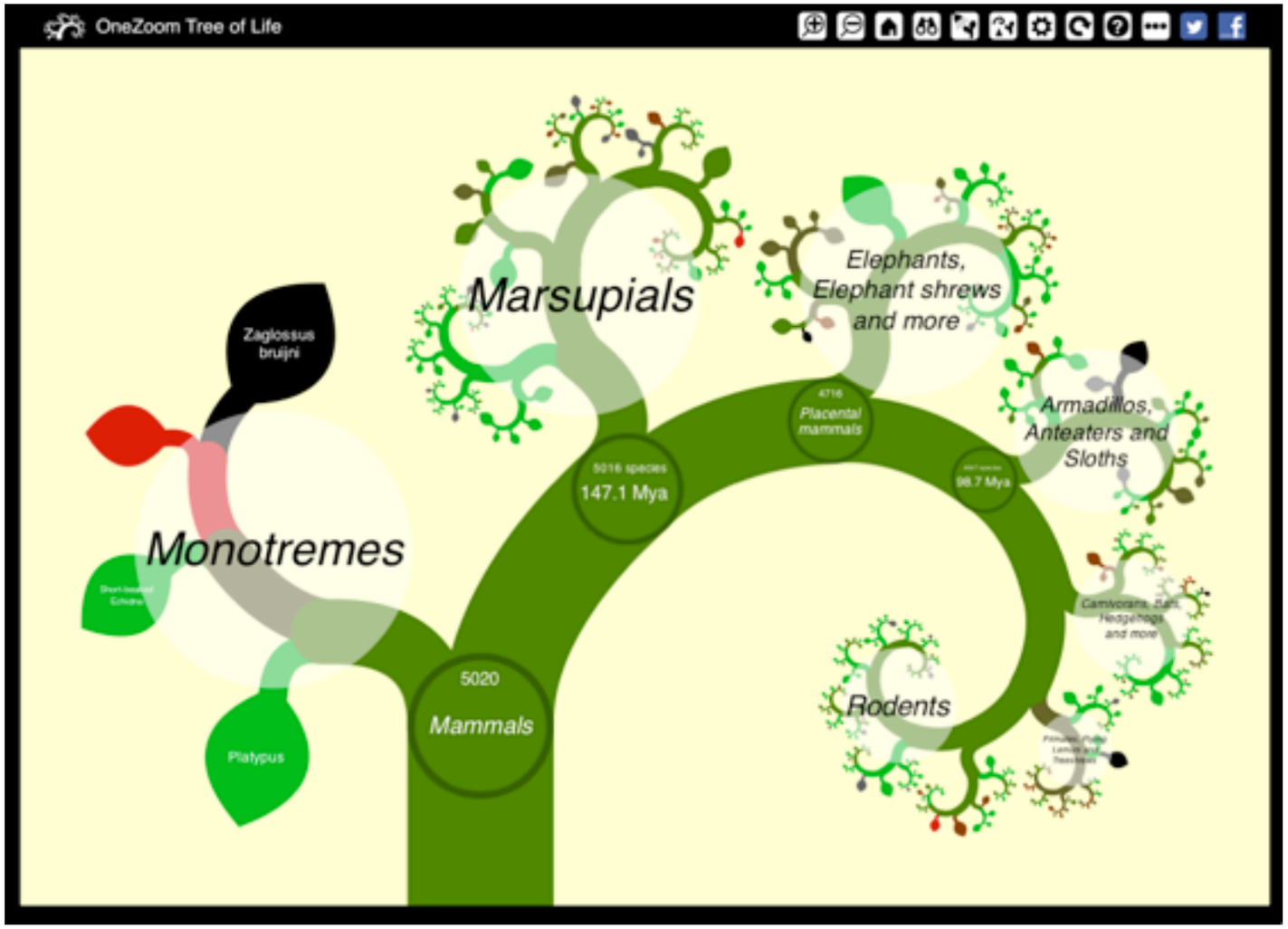


Although Newick may be seen as a kind of standard, there is in reality quite a bit of diversity in interpretation by software and there are a number of variants such as NHX (New Hampshire Extended) and NEXUS, with their XML derivatives phyloXML and NeXML

There are some important differences between file trees and phylogenetic trees. For example, computer file trees commonly have folders which contain just one folder, whereas phylogenetic trees typically have bifurcating or multifurcating nodes (a single parent with 2 or more descendants)

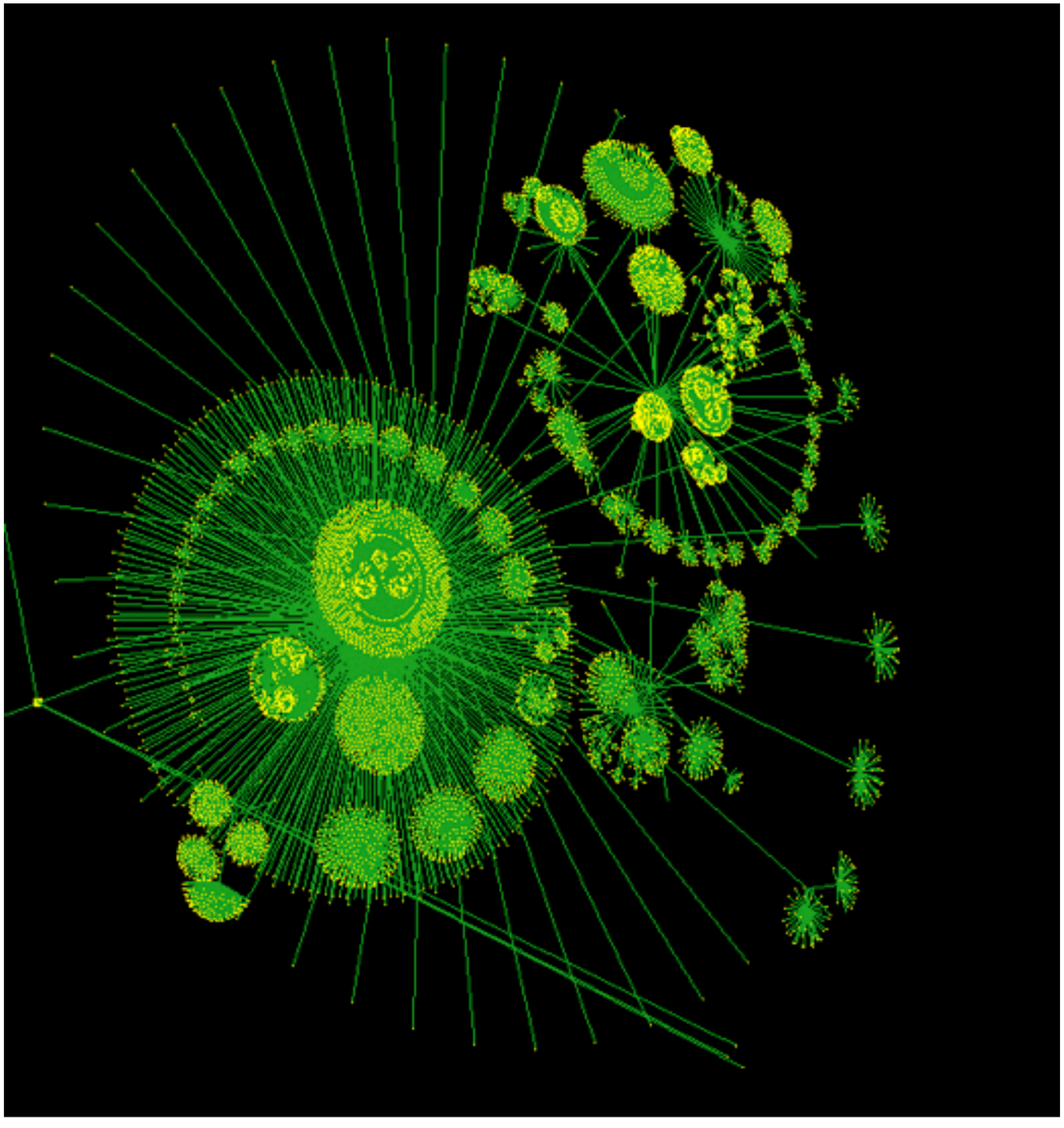
OneZoom Explorer of Imperial College

Recently fractal algorithms have been adopted by OneZoom Explorer enabling interactive visualisation of the tree of life. One manifestation has been a multitouch table connected to a multiscreen system



The eMSS Lab at the British Library and the OneZoom originator James Rosindell have initiated a collaboration to modify, test and demonstrate OneZoom for use with computer directories

3D Tree Visualisation

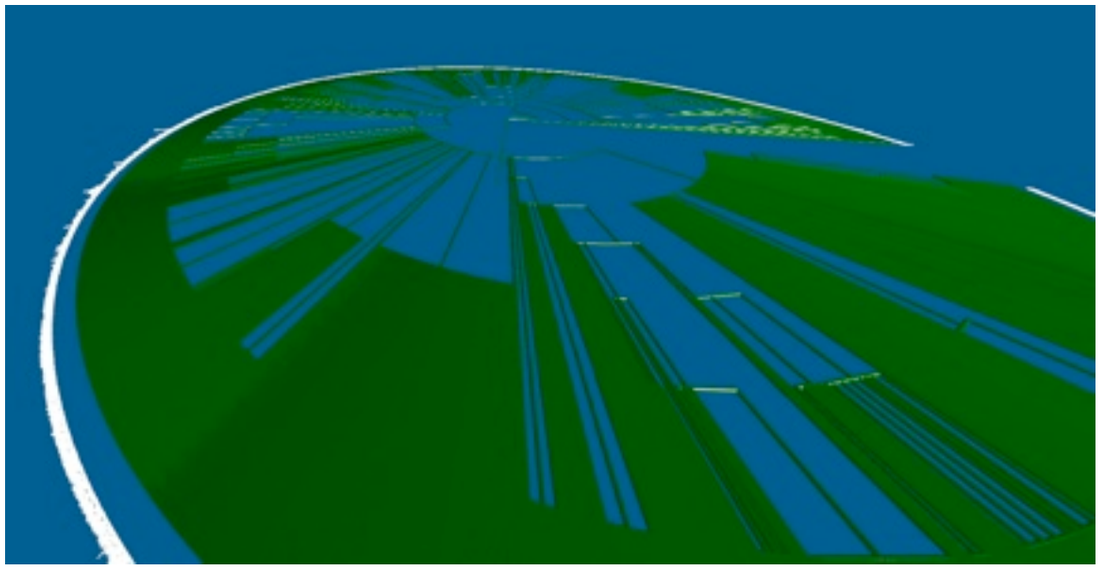


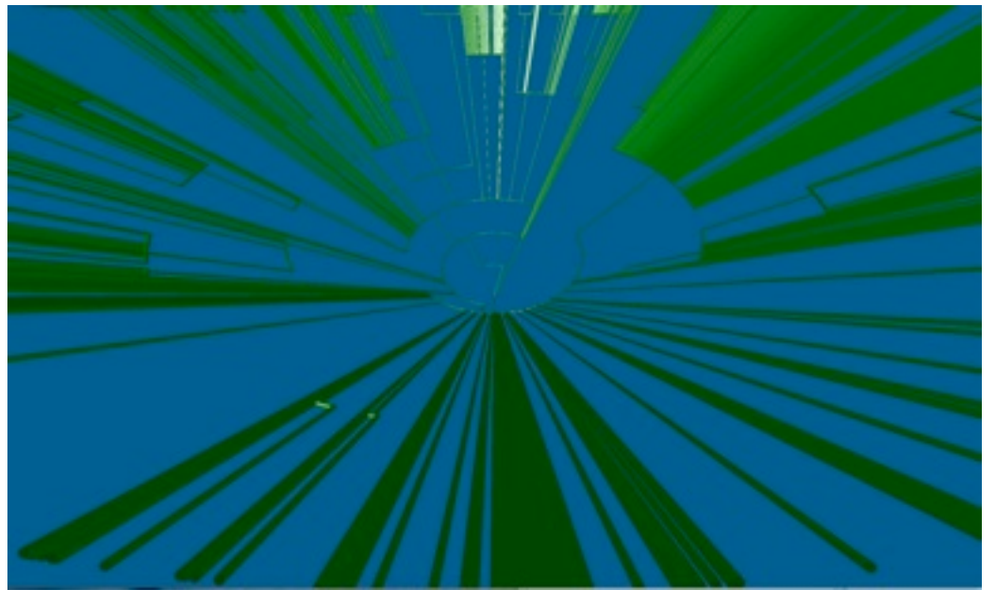
It is still early days for 3D phylogenetic trees and so far the emerging possibilities have had an ambivalent reception but there have been some important efforts.

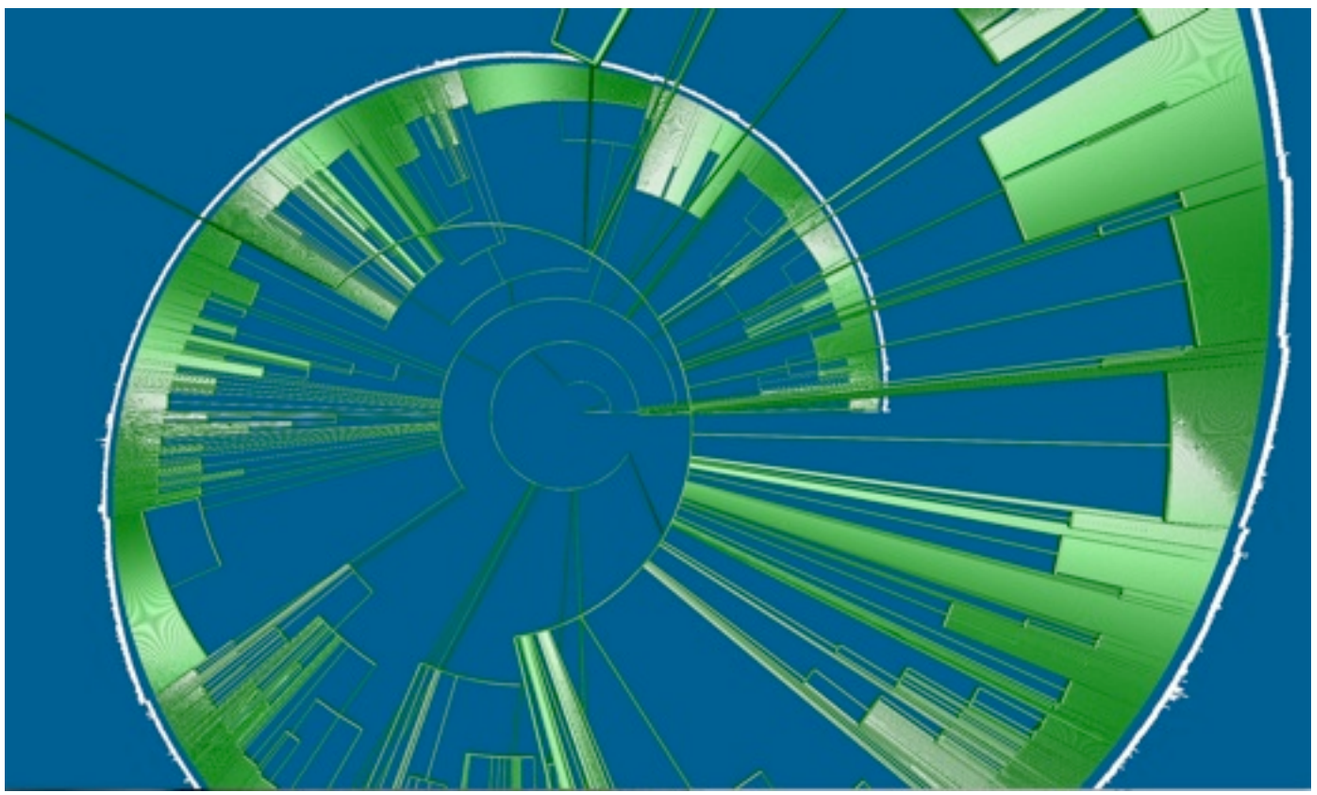
Among the most notable are Paloverde and Phylo3D (which makes it possible to use Walrus)

A screenshot of a file directory tree from a hard drive of John Maynard Smith using Walrus with Phylo3D

3D Tree Visualisation







Three screenshots of the visualisations of a hard drive created using Paloverde: circle, cone, spiral

TreeCurator

- Take disk folder arrangement and create standard graph file
- Create Newick file
- Create standard graph file from Newick file
- Accept standard graph file
- Cope with monotomous nodes and polytomous nodes
- Create bifurcating version of multifurcating tree
- Cope with variants (eg NHX & Nexus and XML and counterparts PhyloXML & NeXML)

Digital Forensics and Preservation

Jeremy Leighton John

DPC Technology Watch Report 12-03 November 2012

Series editors on behalf of the DPC
Charles Beagrie Ltd.

Principal Investigator for the Series
Neil Beagrie

DigitalPreservationCoalition

DPC Technology Watch Series