

CODIL as a knowledge base system for handling historical information

Chris F. Reynolds*

30.1 Introduction

Carrying out any historical research is like trying to solve a jigsaw puzzle in which the vast majority of the pieces are missing. Those that remain may be misunderstood, distorted with time, or highly biased. The possibilities for uncertainty, ambiguity, and error are legion. If only computers could help...

Of course computers can help historians in many ways—but the one area where they are particularly weak is in providing a data base facility for handling the untidy mass of information that is the best we can assemble from the surviving documentary and archaeological evidence. This should not be surprising. Conventional data base techniques are specifically designed to handle large quantities of homogeneous data—because that is what industry, commerce and research scientists want. As a result each record contains a fixed number of predefined fields, most if not all of which contain data. These can be used for certain highly organized and comparatively modern historical records, such as the Victorian census returns, but become less and less suitable as the variety of the information is increased.

At the other extreme there are text-based systems, which can look very impressive to the naive, but have little real information processing capability. This is because there is no record structure and to do anything useful the computer would need a virtually human level of ability to understand (rather than reproduce) the language of the original document. (The examples given in artificial intelligence studies using natural language data bases are typically confined to very small, and hence highly selective, collections of data.)

To overcome these problems many libraries and museums have designed tagged field record systems. These allow variable information to be held about different kinds of artifacts. Each record can consist of a large number of fields, most of which are absent. While ideal for cataloguing and retrieval they are not capable of providing the historian with a sufficiently flexible tool to solve his jigsaw puzzle.

To a historian user, the CODIL and MicroCODIL software, described here, can be considered as a significant extension of the tagged approach. However they incorporate

* CODIL Language Systems
33 Buckingham Road
Tring
Herts

ideas from conventional data bases, artificial intelligence research, and human factor studies, which makes them far more flexible and far more friendly. (CODIL is the original mainframe package, MicroCODIL is more user friendly micro based subset. For discussion purposes they will be assumed to be equivalent.) Their novel features are best introduced by means of Fig. 30.1, based on a 'file dump' of the biography of a George Washington Gibbs.

The first thing to notice about this computerized biography is that you can read it without knowing anything about CODIL or computers. The only difficulty some people have is with the level numbers on the left. Basically an item's influence extends from the point at which it occurs until another item is reached with the same, or a lower, level number. This means that the whole of the listing concerns George, because his name is the only item at level 1. In the same way events between 1832 and 1865 all involve the 'Liverpool' item at level 2.

The second thing to notice is that the information comes from a wide range of historical records in a wide variety of forms. For instance an engraving was made in 1826 which shows the front of George's shop. In the original engraving the words on the boards can be clearly seen, giving the shop keeper's name and occupations. Other information comes from an Aylesbury newspaper (published by George's nephew)—including the reference to a letter which first appeared in the Liverpool Mercury.

The listing shows a number of 'records' or 'statements' describing events in George's life. Each statement consists of a number of items, starting at level 1 and terminated by a full stop. For example one of these statements contains the following items:

```

NAME = Gibbs, George Washington,
PLACE = St Albans, Herts,
YEAR = 1826,
ADDRESS = Clock House,
OCCUPATION = Auctioneer,
PARTNER = Gibbs, John,
SOURCE = Trade Directory.

```

Conventional data bases are restricted to records which contain predefined fields in a predefined order. It is only necessary to look at the full example to realise that such a strait-jacket is totally unacceptable for anyone interested in holding any *realistic representation* of a person's biography. CODIL allows statements to contain any combination of items in any order, including multiple entries—as George's occupations in 1826 illustrate. There are also no constraints on the order of statements in a file. New item names can be created if new kinds of information are discovered—for example I still do not know his salary when he worked for the customs service. Should I find out it would only take a moment to create a SALARY item. Where something happened at a somewhat uncertain date (see his daughters birth dates) it is a routine operation to record a range, as opposed to a single value or a list of values.

The statement relating to 1853 shows another interesting feature. At about this time the Bucks Advertiser carried a regular 'Liverpool' column. It not known who the reporter was. However George was a relative of the editor and he was probably the reporter concerned. Considerable care is needed in handling such uncertain data in a conventional data base—as otherwise you may look at the data several years later and

30. CODIL AS A KNOWLEDGE BASE SYSTEM FOR HANDLING HISTORICAL INFORMATION

1 NAME - Gibbs, George Washington,
 2 YEAR - 1802,
 3 PLACE - Aylesbury, Bucks,
 4 EVENT - Birth,
 5 FATHER - Gibbs, Robert,
 6 MOTHER - Miller, Ann.
 2 YEAR - 1805,
 3 EVENT - Birth Registration,
 4 SOURCE - Dr Williams Library.
 2 YEAR - 1824,
 3 PLACE - St Albans, Herts,
 4 ADDRESS - Clock House,
 5 LANDLORD - St Albans Corporation,
 6 SOURCE - Lease.
 2 YEAR - 1825,
 3 PLACE - Watford, Herts,
 4 EVENT - Marriage,
 5 WIFE - Dagley, Ann.
 2 PLACE - St Albans, Herts,
 3 YEAR - 1826,
 4 ADDRESS - Clock House,
 5 OCCUPATION - Printer,
 6 OCCUPATION - Broker,
 7 OCCUPATION - Pawnbroker,
 8 SOURCE - Antique print.
 5 OCCUPATION - Auctioneer,
 6 PARTNER - Gibbs, John,
 7 SOURCE - Trade directory.
 3 YEAR - 1828,
 4 YEAR - 1829,
 5 DAUGHTER - Gibbs, Ann,
 6 SOURCE - 1851 census.
 2 PLACE - Liverpool, Lancs,
 3 YEAR - 1832,
 4 YEAR - 1833,
 5 DAUGHTER - Gibbs, Eliza,
 6 SOURCE - 1851 census.
 3 OCCUPATION - Landing waiter,
 4 SOURCE - Trade directory,
 5 YEAR - 1835,
 6 YEAR - 1837,
 7 ADDRESS - 13 Hunter Street.
 3 YEAR - 1837,
 4 ADDRESS - Hunter Street,
 5 AUTHOROF - Letter re Lord Nugent,
 6 SOURCE - Liverpool Mercury,
 7 SOURCE - Aylesbury News.
 3 OCCUPATION - H M Customs,
 4 SOURCE - Trade directory,
 5 YEAR - 1841,
 6 ADDRESS - 17 Gildart Street.
 5 YEAR - 1851,
 6 ADDRESS - 33 Netherfield Road North.
 4 SOURCE - Census,
 5 YEAR - 1851,
 6 ADDRESS - 33 Netherfield Road North,
 7 STATUS - Widower.
 3 OCCUPATION (.5) - Journalist,
 4 SOURCE - Bucks Advertiser,
 5 YEAR - 1853,
 6 NOTE - Paper's Liverpool column,
 7 CHECK - Identity not confirmed.
 3 OCCUPATION - H M Customs,
 4 ADDRESS - 33 Netherfield Road North,
 5 SOURCE - Trade directory,
 6 YEAR - 1857.
 4 OCCUPATION - Retired,
 5 YEAR - 1865,
 6 EVENT - Death.
 2 YEAR - 1865,
 3 EVENT - Probate,
 4 PROBATEOFFICE - Principal Registry,
 5 ESTATE 2000,
 6 EXECUTOR - Gibbs, Richard,
 7 EXECUTOR - Gibbs, Robert.

Figure 30.1: George Washington Gibbs of St Albans & Liverpool

forget that it was only an informed guess. As can be seen, CODIL allows a probability to be associated with an item to indicate the reliability of the information.

It is one thing to store historical data in a flexible manner. However it is only useful if it can be processed. I am sure that most historians would recoil in horror at the idea of writing a program which reads records from a file in which every record has a different structure, and where each field may contain a single value, a range or a list. CODIL gets round this problem by using some advance artificial intelligence style techniques and a considerable degree of computer human factors research has gone into the design of MicroCODIL. It is not possible to describe these techniques in detail in this paper but in most cases CODIL does what you might expect a human to do in the same circumstances. For instance a search for 'OCCUPATION = Broker' will find the statement relating to the picture of the shop without the questioner having to know that a person may have more than one occupation simultaneously. A search for events in 1829 will indicate that he had a daughter called Ann Gibbs, although there is some uncertainty about whether she was born in 1828 or 1829.

There are other features which are particularly relevant to the historian. The biography contains references to other people, with roles such as FATHER, PARTNER and EXECUTOR. By including definitions such as:

```
FATHER (ISA) PERSON
MOTHER (ISA) PERSON
...
EXECUTOR (ISA) PERSON
```

it is possible to carry out a search for 'PERSON = Gibbs, John' without knowing the nature of George's link to John when the question is asked. It is also possible to use 'fuzzy' definitions using probability. To take a simple example you might want to define the word PUBLISHING to be any of the following items:

```
OCCUPATION = Publisher
OCCUPATION = Journalist
OCCUPATION (.75) = Printer
AUTHOROF
```

Note that the 'Printer' item has a probability associated with it because some printers do work which is not normally considered as publishing. A simple search for things that are definitely known about PUBLISHING will only find the letter George wrote to the Liverpool Mercury. A search using a 0.6 probability threshold will reveal his printing activities in St Albans. Dropping the threshold to 0.4 allows his possible journalistic activities of 1853 to be reported as well.

30.2 Example studies

Of course, it is one thing to demonstrate flexible handling of historical information, and another to do it on a scale useful to a working historian. For this reason a number of historical studies have been carried out using CODIL on a mainframe computer. The volume of data involved is about four million bytes of compressed data, or over 1000 pages of biographies. The main studies involved are:

1. *The Gibbs Family of Aylesbury*: This involves an in-depth study of this non-conformist family of shop keepers, mainly in the early 19th century. The book, 'The Gibbs Family of Aylesbury' was published using the information from the data base. The biography of George Washington Gibbs is taken from an updated version of the file, modified for use with MicroCODIL. The longest biography, involving over 200 separate statements, describes the career of the politically active John Gibbs.
2. *The Farming Families of Sandridge*: This is an incomplete study, which combines census material with many other sources, which attempts to investigate cousin marriages and migration among well-to-do tenant farmers. Some of the information gathered for this study has been used to set up and demonstrate a local history data base for schools, using MicroCODIL.
3. *The Phipson Family*: An in-depth study has been carried out on this West Midland family. The senior branch moved to Birmingham in the 18th century and become actively involved in manufacturing. They had contacts with Joseph Priestly and Mathew Bolton and married into the Ryland family. Side branches became involved in the medicine, the law and architecture. Another branch were based in the Stourbridge area, and saw the other side of the industrial revolution as nail makers on Lye Waste.

It is appropriate to look at the Phipson files in more detail, to demonstrate that CODIL can do much more than just store and retrieve information. Fig. 30.2 shows the biography of William Howell Phipson listed out a partially formatted manner which takes up less space than a file print. It should be noted that the main participants are all given reference numbers and that these are replaced by the name / birth year before printing. This information can be reorganized in a variety of ways. The simplest is to produce alphabetic indexes, for example by place or occupation. Such indexes not only provide a quick way of means of accessing the information, but also allow aberrant spellings and errors to be spotted and corrected. Summary tables can also be produced and an example is given in Fig. 30.3. Check lists of this nature are extremely useful aids, and it is often possible to produce search aids with the information in an order that corresponds to the manuscripts, etc., to be examined in a records office. Other aids can be used to help with data input, while family trees such as Fig. 30.4 can be 'sideways printed'.

30.3 Technical information

The above discussion, plus the demonstration and sample listings on display at the Conference, show some of the ways that the software can be used in historical research. However the techniques can be used by anyone who has poorly structured information, from an amorphous collection of sources, and, for example, the techniques have been used to handle medical research information. CODIL started with some ideas put forward about 20 years ago. These have been developed over the intervening period at both the software and application level. The existing mainframe software is written in COBOL 76 and is available to other universities, etc. It supports the large studies describe above, and if installed with slight modification on a suitable mainframe it should

- 1 Phipson, Joseph [1745]
 - 2 wife Howell, Anne [1750??]
 - 2 son Phipson, William Howell [1776]
 - 3 wife Evans, Sarah [1786/7]
 - 3 son Phipson, Andrew Bond [1821] ---
 - 3 daughter Phipson, Anne Elizabeth [1823]
 - 4 husband Keeling, Allin [1820??]
 - 4 son Keeling, William Phipson [1848]
 - 4 daughter Keeling, Laura Louisa Sarah [1851??]
 - 5 daughter GUESS Keeling, Laura [1889]
 - 4 son Keeling, Alfred Henry [1860]
 - 4 son Keeling, Ernest Melbourne [1862]
 - 5 wife Cathcart, Catherine
 - 5 daughter GUESS Keeling, Laura [1889]
 - 5 son Keeling, Ernest
 - 5 daughter Keeling, Florence John
 - 3 son Phipson, Alfred William [1826]
 - 4 wife Voice, Mary Louisa [1826??]
 - 4 son GUESS Phipson, Alfred George [1860]
 - 4 wife ?(Phipson), Sarah Amelia Louisa [1830??]
- 2 son Phipson, Joseph [1778] ---
- 2 son Phipson, Thomas [1781] ---
- 2 daughter Phipson, Ann [1783]
 - 3 husband Bond, John Culcope [1775/6]
 - 3 son Bond, John Fletcher [1806]
 - 3 daughter Bond, Rosalinda [1808]
 - 4 husband Van Wart, William
 - 4 son Van Wart, Washington Irving
 - 3 son Bond, William [1810]
 - 3 son Bond, Thomas Phipson [1812]
 - 3 son Bond, Edmond Culcope [1814]
 - 3 son Bond, Samuel Rogers [1814]
 - 3 son Bond, Edmond Culcope [1816]
 - 3 son Bond, Samuel Rogers [1818]
 - 3 son Bond, Henry [1821]
 - 3 daughter Bond, Ann Lydia [1822]
 - 3 daughter Bond, Emily [1825]
 - 4 husband Barber, Nathaniel
- 2 daughter Phipson, Mary [1786]
 - 3 husband Selby, William [1790??]
- 2 wife Floyer, Lydia [1760??]

Figure 30.2

30. CODIL AS A KNOWLEDGE BASE SYSTEM FOR HANDLING HISTORICAL INFORMATION

Phipson, William Howell [1776] ref. no. 26004

1776.07.01 Birmingham sex = m; address = St Phillip; father = Phipson, Joseph [1745]

1776.07.01 Birmingham address = St Phillip; mother = Howell, Anne [1750??]

1776.07.27 Birmingham event = baptism; church = Old Meeting

1788.10.02 (dr Williams Library) witness = Phipson, Joseph; witness = Ryland, Elizabeth; witness = Radnall, Elizabeth; witness = Scholifield, Rev (baptism); birth ref = 2841

1817 note = Sent letter number 6 to Sarah Evans (first dated surviving letter)

1818.09.15 Cardiff note = Wrote to Thomas Evans asking for Sarah's hand

1819.01.11 London event = marriage; church = St George, Hanover Sq; witness = Evans, Thomas; witness = Phipson, Thomas; source = marriage certificate; wife = Evans, Sarah [1786/7]

1821 Landaff note = son baptised

1821.01.20 Cardiff son = Phipson, Andrew Bond [1821]

1823 Cardiff daughter = Phipson, Anne Elizabeth [1823]

1823.11.04 Cardiff address = Fairwater; note = He started a correspondence with Rosalinda Bond in verse

1826.04.02 Cardiff son = Phipson, Alfred William [1826]

1828 directory Birmingham occupation = eating house keeper; address = 6 Moor St; text = Phipson, William

1836.10.25 Birmingham event = bankrupt; occupation = coffee house keeper

1842 directory Birmingham occupation = teacher; address = Day School, Balsall Heath Rd; text = william howard phipson

1843.11.19 Birmingham event = death

1843.11.24 Birmingham event = burial; church = Old Meeting; age = 67; text = william howell phipson, late of balsall heath road, buried in grave no 31, new ground

Figure 30.3

Ref no.	Name	Birth	Bapt	B Reg	Marr	M Reg	Died	Buried	D Reg	Prob
61138	Phipson, William [1617]	1617	1617							
32890	Phipson, William [1678]	1678	1678							
59664	Phipson, William [1725??]	1725??								
58839	Phipson, William [1753]	1753	1753							
27720	Phipson, William [1769]	1769	1769							
35288	Phipson, William [1797]	1797	1797			1798??		1845		1845
58806	Phipson, William [1810??]	1810??				1820		1874		1874
35354	Phipson, William [1824]	1824	1824			1835??				
62535	Phipson, William [1830??]	1830??				1846	1846	1908		1908
65835	Phipson, William [1849]	1849				1855??				
35706	Phipson, William [1859]	1859	1859	1859				1849	1849	
35849	Phipson, William [1861]	1861			1861			1925		1925
34628	Phipson, William [1871]	1871						1862		1862
49874	Phipson, William [1880??]	1880??	1881					1871		1871
34474	Phipson, William [1893]	1893	1893	1893		1903	1903	1916		1916
36542	Phipson, William [1901]	1901			1893			1895		1895
29535	Phipson, William [1915]	1915			1901	1929	1929	1980		1980
26048	Phipson, William Andrew [1847]	1847			1915			1915		1915
49797	Phipson, William Andrew [1901]	1847			1847	1873	1873	1885	1885	1885
49929	Phipson, William Edward Gordon [1937]	1901								
26004	Phipson, William Howell [1776]	1937				1960				
49885	Phipson-Errington-Yates, William L_____ [1915??]	1776	1776	1788	1819			1843	1843	
28083	Phipson, William Maynard [1847]	1915??						1942		1942
27786	Phipson, William Price [1806]	1847				1876	1876	1908		1908
27984	Phipson, Wilson Weatherley [1838]	1806	1807					1839		1839
		1838	1838			1867	1867	1891		1891

Figure 30.4

be able to support files of 10 million bytes or more without difficulty (unfortunately the computer on which it runs at Brunel University was withdrawn in June 1988 so that further work there has had to be stopped.) MicroCODIL is a version of CODIL designed to run on a BBC microcomputer (any model) in a school environment. It is aimed at showing how a wide range of 'information technology' ideas can be used in 'across the curriculum' studies, especially history. For instance there is a special 'History Project Pack' which includes a copy of the book, 'MicroCODIL and History'. MicroCODIL has a much better user interface than CODIL, and has some powerful extra facilities, such as the use of probability, described above. However the small size of the BBC computer means that its file handling capabilities are much less. Future plans are to move this system to larger computers, rather than to upgrade the older CODIL package.

Reynolds 1987 and Reynolds 1989 describe the basic concepts underlying the CODIL / MicroCODIL approach, and replaces earlier papers on how the software works. Reynolds 1984 and Reynolds 1985 provide descriptions of the prototype version of MicroCODIL as a teaching package. Reynolds 1988 is based on the package currently being marketed. Reynolds 1979 consists of computer printed biographies and family trees, with some connecting text. The MicroCODIL software and documentation, including the 'MicroCODIL Manual' and 'MicroCODIL and History', are available from the author at the CODIL Language Systems address.

References

- REYNOLDS, C. F. 1979. *The Gibbs Family of Aylesbury*.
- REYNOLDS, C. F. 1984. "MicroCODIL as an Information Technology Teaching Tool", *University Computing*, 6: 71-75.
- REYNOLDS, C. F. 1985. "A Microcomputer Package for Demonstrating Information Processing Concepts", *J. Microcomputer Applications*, 8: 1-14.
- REYNOLDS, C. F. 1987. "Human Factors in System Design—A Case Study". in Diaper & Winder, (eds.), *People and Computers III*, pp. 93-102. Cambridge University Press.
- REYNOLDS, C. F. 1988. "Introducing Expert Systems to Pupils", *J. Computer Aided Learning*. in press.
- REYNOLDS, C. F. 1989. "CODIL—The Architecture of an Information Language", *Computer Journal*.