The International Journal of Digital Curation

Issue 3, Volume 4 | 2009

Curating the CIA World Factbook

Peter Buneman, Heiko Müller,

School of Informatics, University of Edinburgh

Chris Rusbridge,

Digital Curation Centre, University of Edinburgh

Abstract

The *CIA World Factbook* is a prime example of a *curated database* – a database that is constructed and maintained with a great deal of human effort in collecting, verifying, and annotating data. Preservation of old versions of the Factbook is important for verification of citations; it is also essential for anyone interested in the *history* of the data such as demographic change. Although the Factbook has been published, both physically and electronically, only for the past 30 years, we appear in danger of losing this history. This paper investigates the issues involved in capturing the history of an evolving database and its application to the CIA World Factbook. In particular it shows that there is substantial added value to be gained by preserving databases in such a way that questions about the change in data, (longitudinal queries) can be readily answered. Within this paper, we describe techniques for recording change in a curated database and we describe novel techniques for querying the change. Using the example of this archived curated database, we discuss the extent to which the accepted practices and terminology of archiving, curation and digital preservation apply to this important class of digital artefacts.[1]

[1] This paper is based on the paper given by the authors at the 5th International Digital Curation Conference, December 2009; received November 2009, published December 2009.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The term “curation” comes from the Latin *curare* – to care for. *Curated databases* are databases that are populated and updated with a great deal of human effort through the consultation, verification, and aggregation of existing sources, and the interpretation of new raw data (Buneman, Cheney, Tan & Vansummeren, 2008). Many curated databases are published online, largely replacing the material – dictionaries, encyclopedias, gazetteers and so on – that one traditionally found on the reference shelves of libraries. Moreover, because the Internet has made it relatively easy to publish such information, there has been an explosion in the number of curated databases, especially in the sciences. The field of molecular biology, for example, boasts of over 1,000 online databases, all of which, to some extent, are curated (Galperin, 2008). As artefacts, curated databases can be expensive to produce. One of the leading molecular biology databases, UniProt[2] is maintained by over 100 staff members (Schneider, Bairoch, Wu & Apweiler, 2005) and others, such as the IUPHAR database[3], while only employing one or two full-time curators, are dependent on hundreds of “contributors” – senior researchers who voluntarily put substantial effort into providing content for the database. While curated databases require substantial effort in creating the database and maintaining its accuracy, it appears that very little effort has been put into curating such databases, in the sense of preserving their history and enhancing their long-term usefulness.

We take a broad view of what is meant by a database to include, for example, ontologies and structured files. For our purposes, the distinguishing properties are first that there is some internal structure, often some kind of hierarchical format, and second that the contents – and possibly the structure – evolve over time. Taking these two terms together, the CIA World Factbook (Central Intelligence Agency [CIA], 2009) is an excellent example of a curated database. It is almost certainly the most highly used online reference for demographic information. While there may be questions about the provenance of some of its estimates, the fact that it is so widely used makes it a prime candidate for preservation. Like many other reference sources, the Factbook is constantly updated. However, changes to the Factbook data are rarely documented and now the online version is overwritten whenever changes occur. Although the Factbook has only been published, initially in print and more recently electronically, for the past 30 years, we are in danger of losing its history. Moreover, as the Factbook evolves, it is becoming increasingly difficult to bring the historical data into a unified and usable format. Therefore, one would like not only to preserve all versions of the Factbook, but to preserve them in such a way that trends or “longitudinal” information such as “How did the GDP of Lithuania change in the past 10 years?” or “When was the Czech Republic created?” can be easily extracted. Within this paper, we describe our ongoing efforts in archiving the CIA World Factbook, focusing on the preservation of the database history in a usable form.

The goals of this paper are threefold. First, it demonstrates a method we have developed for archiving the history of an evolving database. The method provided allows the retrieval of the database as it was at some past time (often called a snapshot). Secondly, it describes a new query system that facilitates the use of temporal or longitudinal queries such as those described above. Finally, there is a

[2] UniProt http://www.uniprot.org

[3] IUPHAR Committee on Receptor Nomenclature and Drug Classification http://www.iuphar-db.org

whole methodology and terminology that has been developed in connection with archiving digital artefacts. We examine the extent to which it can be meaningfully applied to curated databases. The title of this paper claims that we are *curating* the Factbook – a collection of structured documents that have, individually, already been curated! We believe this to be an appropriate use of the term. According to the Digital Curation Centre,[4] digital curation is "... maintaining and adding value to a trusted body of digital information ...". By facilitating longitudinal views of the data, we are adding value, and – as we shall argue in this paper – we believe that we are providing a technology for the long-term maintenance of evolving databases.

The paper is structured as follows. We first describe the background, how the Factbook has evolved and is evolving, and how we collected – and continue to collect – versions of it. Second, we describe how the versions are efficiently combined into a single archive that can be used for both snapshot and longitudinal queries. Finally we compare what we are doing here with the accepted processes of digital curation for fixed objects.

Finding the Editions

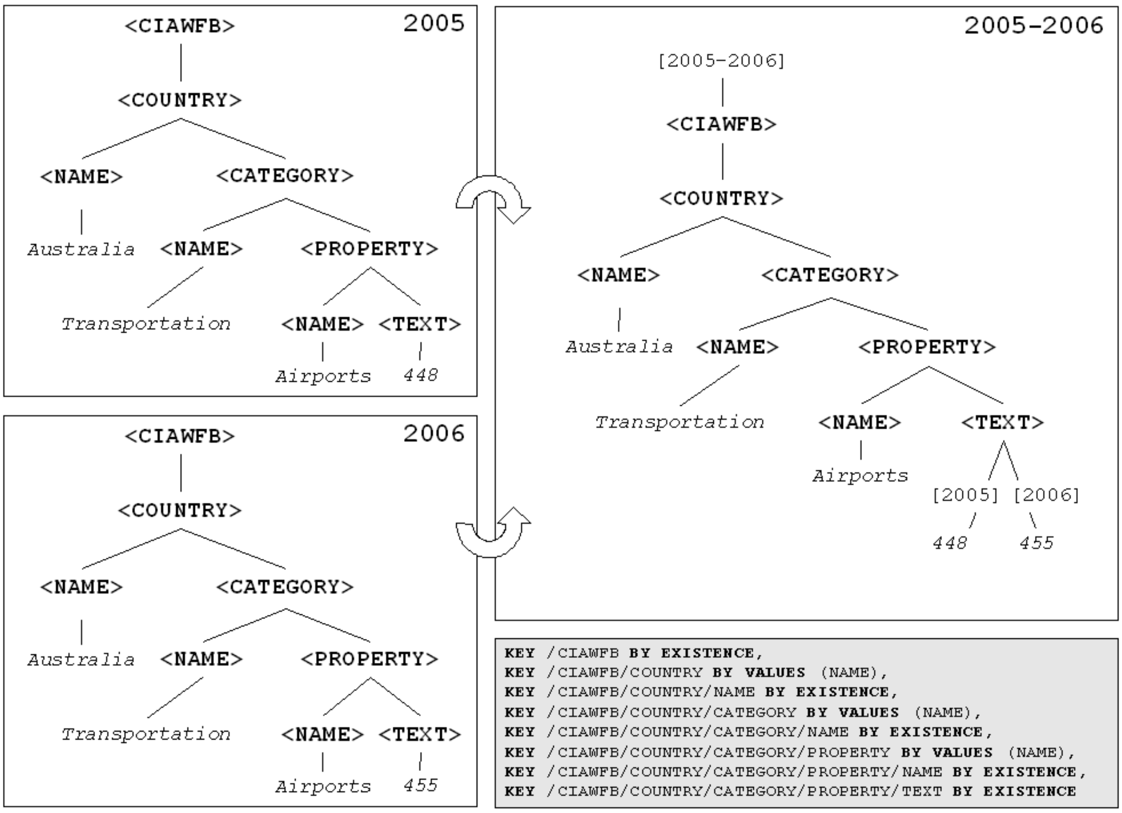
The Factbook was created as an annual summary and update to the encyclopedic National Intelligence Survey studies. The first classified Factbook was published in August 1962, and the first unclassified version was published in June 1971. The 1975 Factbook was the first to be made available to the public by the US Government Printing Office. The National Basic Intelligence Factbook was produced semiannually until 1980. Starting in 1981, publication became an annual product and was renamed The World Factbook. The Factbook was first made available on the Internet in June 1997[5]. The CIA itself provides online editions for 2000 onwards, and states that "Hard copy editions for earlier years are available from libraries". Starting in 2008 the online edition has been updated whenever deemed appropriate (currently bi-weekly). Recently the CIA announced that it will no longer offer the Factbook in printed editions, although the 2008 and 2009 editions are being published by the US Government Printing Office.

Assembling the complete history poses an additional challenge for preservation. In fact, we have been able to find all editions from 1990 to 2000 on the Web. Editions prior to 1990 are currently not known to exist in digital form, but only in print. Editions 1990-1994 are "plain vanilla" text and have been transcribed by Project Gutenberg[6]. Editions 1995-2000 are in HTML and have been archived by The University of Missouri-St. Louis[7]. However, the 1999 edition is corrupted, and finding an uncorrupted version was not easy. Its survival in a forgotten archive appears accidental. We found it among editions 1998-2001, available from The University of Würzburg in the "Elwis Mirror" – an organization that has almost no Internet presence and now appears to be defunct[8].

[4] What is Digital Curation? http://www.dcc.ac.uk/about/what
[5] The World Factbook https://www.cia.gov/library/publications/the-world-factbook/docs/history.html
[6] Project Gutenberg http://www.gutenberg.org/wiki/Main_Page
[7] University Libraries of the University of Missouri http://www.umsl.edu/services/govdocs/
[8] Formerly at http://www.wifak.uni-wuerzburg.de/fact99

Combining the Editions

Throughout its (electronic) history, editions of the Factbook have maintained a remarkably consistent structure. It is this structure that allows us to treat all editions, even those originally provided only in text form, as versions of a database. The Factbook contains an entry for each country which is broken down into named *categories, properties*, and (optional) *sub-properties*. For example, the land area of Poland would currently be found under `Country:Poland` → `Category:Geography` → `Property:Area` → `Subproperty:land`. Figure 1 shows another example from the Factbook regarding the total number of airports in Australia in 2005 and 2006. In general, every data element in the hierarchical structure of any edition of the Factbook has a canonical path, and it is this property, which holds for many curated databases, that is required by our archiving technique (described below). Until recently, this property was implicit in the document structure; but in 1997 the CIA started to publish an explicit schema[9]. It also appears from formatting irregularities and other evidence that for most of its history the Factbook has been generated "by hand" and not from some underlying database. Recently such irregularities have gone, suggesting that a database management system is in use.



```
KEY /CIAWFB BY EXISTENCE,
KEY /CIAWFB/COUNTRY BY VALUES (NAME),
KEY /CIAWFB/COUNTRY/NAME BY EXISTENCE,
KEY /CIAWFB/COUNTRY/CATEGORY BY VALUES (NAME),
KEY /CIAWFB/COUNTRY/CATEGORY/NAME BY EXISTENCE,
KEY /CIAWFB/COUNTRY/CATEGORY/PROPERTY BY VALUES (NAME),
KEY /CIAWFB/COUNTRY/CATEGORY/PROPERTY/NAME BY EXISTENCE,
KEY /CIAWFB/COUNTRY/CATEGORY/PROPERTY/TEXT BY EXISTENCE
```

Figure 1. Examples from the Factbook in 2005 and 2006 (left) together with the resulting archive (right-top) and the key specification for the Factbook (right-bottom).

In spite of the general consistency of the documents, a certain amount of data cleaning was needed. We discuss issues of authenticity that can result from data cleaning later. The errors were of two general forms: format, and structure. In some

[9] The World Factbook https://www.cia.gov/library/publications/the-world-factbook/docs/profileguide.html

cases the errors were detected on parsing the documents into the XML format required by the archiver; in other cases they were not detected until we tried to construct the integrated archive. Furthermore, there are still several quality issues about data values in the Factbook, some of which are outlined below.

Format

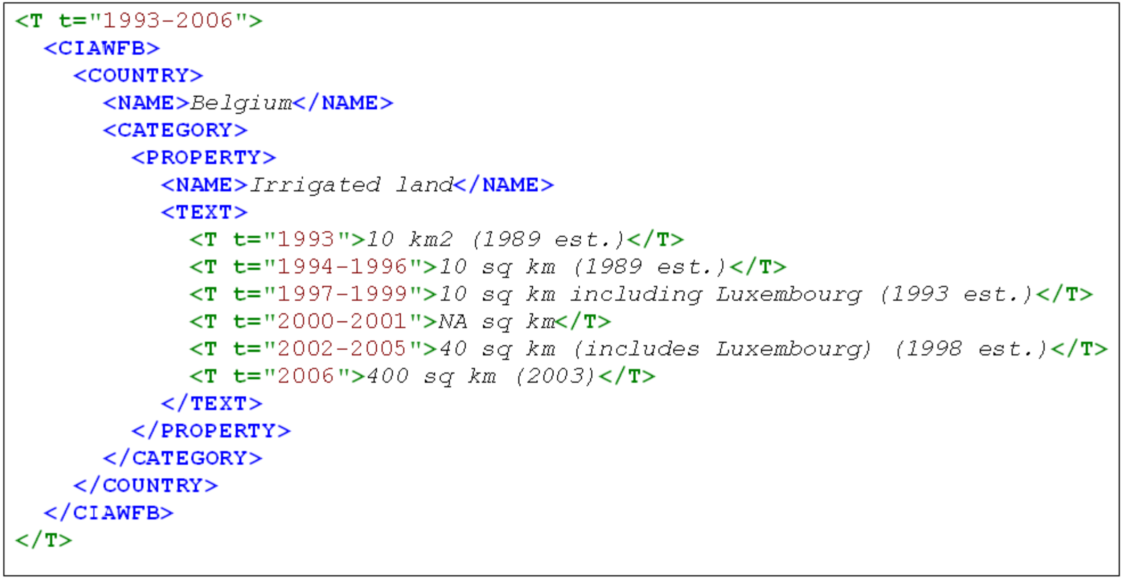
Project Gutenberg claims to keep documents in "plain vanilla" text. However the text had actually been marked up – by a combination of indentation and insertion of non-alphanumeric characters to aid in searching. Each of the 1990-1994 editions adopts a different convention, which required the construction of individual parsers. Unfortunately, the markup, which appears to have been added in the process of manual data entry, had several errors. These were easy to spot (usually in the initial parsing phase) and to correct. In the later HTML versions, there were few such formatting errors. An open problem remains with the introduction of sub-properties in 1992 as explicitly marked-up entities. Until then, many properties contained loosely structured lists of values. Unifying editions 1990-1991 with those from 1992 onwards is currently ongoing work in our archiving effort.

Structure

In one edition, there were a few places in which the document deviated from the "fully keyed" structure (described later) required by our software: there were two sub-properties with the same name. This happened when the country involved was divided, and the curator had found it appropriate to have two sub-properties with the same name rather than a single sub-property with two components. However this policy was not uniform and we chose to deal with this by simply merging the two sub-properties. A more fundamental problem arises when a property was renamed or moved from one category to another between editions (examples later). While this does not create a problem for our software, longitudinal queries on such data are liable to be misleading. For the time being we have not corrected the data in this respect, and how to deal generally with this kind of anomaly is the subject of ongoing research.

Data Quality

We should remark that it is possible that the process of manual transcription by Project Gutenberg also introduced errors. So far we have only detected errors in the formats. We have not yet found any errors in data values (evidence for them would be found in anomalous longitudinal sections), but they may well exist. However, there are several examples of data quality problems in the archived data that have yet to be properly addressed. Figure 2 shows an excerpt from the Factbook that lists the number of square kilometres of land area in Belgium that were artificially supplied with water between 1993 and 2006. The example highlights several data quality issues. First, there is no standardized representation for square kilometres, that is, km^2 vs. sq km. Furthermore, the data values often contain additional meta-information. In particular, additional information about the estimation of particular values raises questions about (a) the accuracy of values, and (b) the timeliness of the data. There are also cases in which data values disappear and reappear between annual releases of the Factbook, thus affecting the completeness of the data. Within our current research we are addressing the problem of applying existing data cleaning techniques to archived data while maintaining both the original and the clean data within a single archive. In general, the process of data cleaning involves a significant amount of work; and our colleagues[10] who have attempted to archive other forms of structured data, have similarly reported on the considerable effort that they put into data cleaning. In retrospect it is surprising that this is not discussed more in digital curation literature.



```
<T t="1993-2006">
  <CIAWFB>
    <COUNTRY>
      <NAME>Belgium</NAME>
      <CATEGORY>
        <PROPERTY>
          <NAME>Irrigated land</NAME>
          <TEXT>
            <T t="1993">10 km2 (1989 est.)</T>
            <T t="1994-1996">10 sq km (1989 est.)</T>
            <T t="1997-1999">10 sq km including Luxembourg (1993 est.)</T>
            <T t="2000-2001">NA sq km</T>
            <T t="2002-2005">40 sq km (includes Luxembourg) (1998 est.)</T>
            <T t="2006">400 sq km (2003)</T>
          </TEXT>
        </PROPERTY>
      </CATEGORY>
    </COUNTRY>
  </CIAWFB>
</T>
```

Figure 2. Data quality issues in the Factbook regarding accuracy, completeness, and timeliness of values exemplified on the number of square kilometres of land area in Belgium that were artificially supplied with water between 1993 and 2006.

The Archiver at Work

Curated databases are predominantly kept in well-organised hierarchical data formats having a key structure that provides a canonical identification for each element by the path in which it occurs and the values of some of its sub-elements. Our archiver uses this property to maintain multiple versions of the same data set efficiently. We distinguish between two different types of nodes in a hierarchy: (i) *element nodes* having a label, and (ii) *text nodes* having a value. Only element nodes may occur as internal nodes. Element keys are defined using key constraints that are similar to keys for XML defined in Buneman, Davidson, Fan, Hara and Tan (2001). A *key specification* K is a set of *key definitions* $k = (q, s)$, where q is an absolute path of element labels and s is a *key value expression*. Each key definition (q, s), specifies a set of elements reachable by path q and defines how the key value is derived from the elements subtree. We distinguish between three types of key value expressions: (i) *existence*, (ii) *subtree*, and (iii) *values*. All element keys are relative keys, that is, the key value uniquely identifies an element among its siblings. Elements that are keyed by existence are keyed by their label. Elements that are keyed by subtree are uniquely identified by the value of their whole subtree (see Buneman, Khanna, Tajima and Tan (2004) for definitions of subtree values). For elements e that are keyed by values, an additional set of relative path expressions $\{p_1, ..., p_k\}$ is given. Each p_i specifies an element in the subtree of e whose value is used as part of the key value for e. These values are referred to as *key path values*. Figure 1 shows (part of) the key specification for the Factbook, stating that countries, categories and properties are identified by their respective NAME element, whereas NAME and TEXT are keyed by existence, that is, they appear only once among the children of their parent.

[10] Kevin Ashley and Wenfei Fan, personal communications

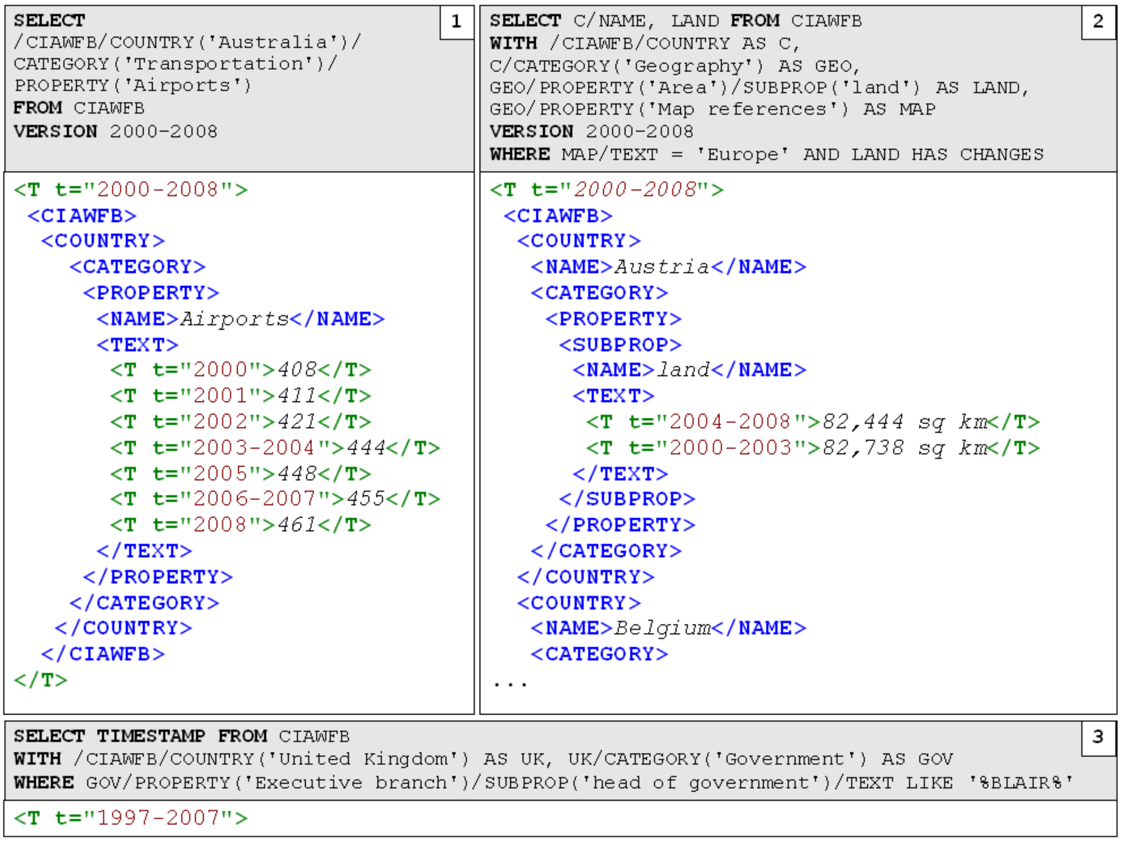
In Buneman et al. (2004) a nested merge approach to archiving is developed that efficiently stores multiple versions of hierarchical data in a compact archive by “pushing down” time and introducing timestamps as an extra attribute of the data. Archives of multiple database versions are generated by merging the versions into a single hierarchical data structure. Corresponding elements in different versions are identified based on their key values. The archiver stores each element only once in the merged hierarchy to reduce storage overhead. Archived element and text nodes are annotated with timestamps representing the sequence of version numbers in which the node appears. Figure 1 shows the resulting archive from merging the two examples from the Factbook in 2005 and 2006. Timestamps are shown in square brackets as edge labels of their respective nodes. Note that we only show (and materialize) the timestamps for those nodes whose timestamp differs from their parent’s timestamp. This nested merge approach has several advantages regarding storage space, retrieval of database versions, and tracking of object history. For example, the 19 annual releases of the Factbook between 1990 and 2008 contain a total of 3,770,468 nodes, whereas the resulting archive contains only 765,355 nodes.

Although the information in the Factbook changes quite significantly in between the years, there is a reduction in storage space from 95.7 MB for the set of individual releases to 39.1 MB for the archive. Even compression does not remove the advantage: if the files are compressed with gzip the reduction is from 19.1 MB to 6.8 MB. When querying archives, the nested merge approach is advantageous over delta-based approaches that maintain records of changes between pairs of consecutive versions. Retrieval of database versions and tracking of object history from archives in delta-based representations may either involve undoing or applying many deltas or require reasoning with the deltas. Retrieval of a version from merged archives, on the other hand, requires a single scan of the data.

Archiving Tool

Based on the algorithms described above, we implemented the archive management system XARCH (Müller, Buneman & Koltsidas, 2008). The system allows one to create new archives, to merge new versions of data into existing archives, and execute both snapshot and temporal queries using a declarative query language. XARCH is currently capable of archiving XML documents as well as data from relational databases. We further provide specific parsers for UniProt flat files, and for the CIA World Factbook HTML pages on the Web. In XARCH we currently use XML as the storage format of archives. Element nodes and timestamps are represented as XML elements. Text nodes are XML strings (see Figure 3 for examples). Our declarative query language XAQL allows retrieval of particular data versions, tracking of object history, and retrieval of timestamps representing the sequence of versions when a given condition was valid. XAQL is oriented towards OQL (Cattell et al, 1997) since archives are not arbitrary XML documents, but follow a fairly regular structure given by the key specification. We consider nodes as objects and an archive as a nested, timestamped object hierarchy. Timestamps become a first class concept in XAQL.

Figure 3 shows three example XAQL queries for the Factbook. The first query shows how the total number of airports in Australia changed between 2000 and 2008. Apart from SQL-like `SELECT`-, `FROM`-, and `WHERE`-clauses, XAQL has a `VERSION`-clause that allows one to restrict the data versions considered in a query. The second query retrieves the name and land area of all European countries with changes to the land area between 2000 and 2008. The predicate `HAS CHANGES` evaluates to true if the specified node or one of its sub-nodes is timestamped. The third query retrieves a timestamp of all Factbook releases that list Tony Blair as Prime Minister of the U.K.



```
SELECT                                                 1
/CIAWFB/COUNTRY('Australia')/
CATEGORY('Transportation')/
PROPERTY('Airports')
FROM CIAWFB
VERSION 2000-2008
```

```
<T t="2000-2008">
 <CIAWFB>
  <COUNTRY>
    <CATEGORY>
     <PROPERTY>
      <NAME>Airports</NAME>
      <TEXT>
       <T t="2000">408</T>
       <T t="2001">411</T>
       <T t="2002">421</T>
       <T t="2003-2004">444</T>
       <T t="2005">448</T>
       <T t="2006-2007">455</T>
       <T t="2008">461</T>
      </TEXT>
     </PROPERTY>
    </CATEGORY>
   </COUNTRY>
  </CIAWFB>
</T>
```

```
SELECT C/NAME, LAND FROM CIAWFB                        2
WITH /CIAWFB/COUNTRY AS C,
C/CATEGORY('Geography') AS GEO,
GEO/PROPERTY('Area')/SUBPROP('land') AS LAND,
GEO/PROPERTY('Map references') AS MAP
VERSION 2000-2008
WHERE MAP/TEXT = 'Europe' AND LAND HAS CHANGES
```

```
<T t="2000-2008">
 <CIAWFB>
  <COUNTRY>
   <NAME>Austria</NAME>
   <CATEGORY>
    <PROPERTY>
     <SUBPROP>
      <NAME>land</NAME>
      <TEXT>
       <T t="2004-2008">82,444 sq km</T>
       <T t="2000-2003">82,738 sq km</T>
      </TEXT>
     </SUBPROP>
    </PROPERTY>
   </CATEGORY>
  </COUNTRY>
  <COUNTRY>
   <NAME>Belgium</NAME>
   <CATEGORY>
...
```

```
SELECT TIMESTAMP FROM CIAWFB                                                     3
WITH /CIAWFB/COUNTRY('United Kingdom') AS UK, UK/CATEGORY('Government') AS GOV
WHERE GOV/PROPERTY('Executive branch')/SUBPROP('head of government')/TEXT LIKE '%BLAIR%'
```

```
<T t="1997-2007">
```

Figure 3. Three temporal queries for the Factbook. Results are shown in XML format.

Problems and Challenges.

When archiving databases over long periods of time, schema changes become an issue. We distinguish between changes (i) to key path values, and (ii) to the key specification. The latter has not been a major issue for the Factbook so far (we have already mentioned the remarkably consistent structure). A few elements have been added and/or removed throughout the years, but the archiver is capable of handling these changes without any problems. If key path values are modified, however, the archiver treats the corresponding elements as distinct elements and does not merge them.

These changes do occur in the Factbook and they limit our ability to track an object's history. For example, the property that lists the names of major ports for a country was called `Ports` from 1990 to 1996, `Ports and harbors` from 1997 to 2005, and `Ports and terminals` since 2006. We are currently developing methods to detect such key path value changes based on complementary timestamps and subtree similarity. Once we are able to detect these changes, we merge corresponding elements in a post-processing step and annotate their key values with appropriate timestamps. More significantly, problems arise when elements are moved between different parents. For example, the total number of airports for each country is found under category `Country:`*`name`* → `Category:Transportation` in Factbooks from 1996 onwards. In 1995, the property was listed under `Country:`*`name`* →

`Category:Economy`, and prior to 1995 under `Country:`*`name`* → `Category: Communications`. Detecting these changes is similar to detecting renamed properties. In order to maintain these changes in a query-able form, however, we would have to extend our data model and consider archives as graphs instead of trees. While merging successive versions of these graphs will work as before, it is questionable whether XML would still be the appropriate storage format for the resulting archive.

We have described the kinds of changes to both structure and contents that have been problematic in our work with the Factbook, but we should also mention changes that cause problems in other databases. In the case of the Factbook, we have seen that a problematic change arises when a country splits or when two countries merge; but this is not a problem for our archiving techniques, it is only a problem for constructing queries that operate on the database. For example if one wants to plot the population of Germany over time, should one attempt to include the period in which Germany was two countries? A more fundamental problem is *object fusion*, which can happen in a variety of databases: here two entries in the database that were thought to represent different real-world entities turn out to represent the same entity. How does one indicate this change in the key structure of the hierarchical data representation, and how does one evaluate longitudinal queries over such objects? The converse of *object fission* – when one database object turns out to represent two real-world entities – is equally challenging.

The second problem not demonstrated by the Factbook is that of more drastic schema changes. For example, we have developed a tool that takes a relational database and, from its schema, creates a key structure and converts the contents of that database into an appropriate XML hierarchy that conforms to the key structure. A change to the relational schema will cause a change to the hierarchical key structure, and while this will not "break" our archiver, it is not all clear how we create useful longitudinal queries across this change in structure. This is very much an area for further research.

General Usefulness

We have described technology for preserving and querying the history of a curated database with special reference to its use with the Factbook. It is both legitimate and important to ask about the general usefulness of this technology, and how many databases resemble the Factbook in their need for this form of preservation. The answer comes in two parts. First, keeping all past states of an evolving curated database is essential for verification. If we cite a database, surely we should cite the version in which we found the relevant information and not the version current at some later time, and one would hope to be able to find that version (Buneman, 2006). For many curated databases, especially the ones we have mentioned in molecular biology, the databases are expected to evolve towards "the truth". While there are some corrections to existing data, most of the changes consist of the addition of data and annotations. For these databases, it is essential to keep previous versions, but longitudinal queries are less important. There are also many databases that record the changing state of the world. These include gazetteers, such as the Factbook, geospatial databases, business intelligence information, census data, and – most importantly – clinical studies. For all of these databases, the ability both to preserve past states and to investigate temporal structure of the data is essential.

Over the years, our archiving technique has been applied to a number of different curated databases. The initial work showed the space-efficiency of archiving the curated databases OMIM[11] and SWISS-PROT[12] in an hierarchical format (Buneman, Khanna, Tajima & Tan, 2004). We extended the nested-merge approach to work on arbitrary large databases, thereby enabling the archiver to maintain databases like the UniProt Knowledgebase[13] where each different version is several GB in size (Koltsidas, Müller & Viglas, 2008). Furthermore, we are currently archiving snapshots of the DBLP Computer Science Bibliography database[14] and the UK's National Weather Service (MetOffice) UK weather observations[15] on a regular basis. The latter is particularly interesting as a highly volatile database. Nevertheless, the archiver compresses the weather data surprisingly well. For example, a collection of 6,000 different snapshots requires 416.08 MB of storage space (14.95 MB when compressed using gzip). An archive of the same data requires only 19.73 MB of storage space, or 8.47 MB when compressed using gzip. Archiving the MetOffice data also allows us to verify that the archiver is capable of efficiently maintaining archives of several thousands of snapshots.

Archiving, Preservation and Curation in this Context

We have so far used the terms archiving, preservation and curation rather loosely. These terms are often used ambiguously, or differently in different communities. It is worth exploring them in the context of curated databases such as this. We start with the term archiving since the main tool in our preservation effort is most often referred to as a database archiver.

Archiving

In a computing context, the most basic meaning of archiving is to store for a period of time, for example, to prevent a temporary file from being lost ("Did you archive that file?"). This form of archiving usually means to make backup copies of the data using different media, for example, hard discs, CDs, tapes and so on. Such copies may be taken on a regular basis ("We archive our data frequently"), but such copies are often subsequently over-written, and are not usually intended to maintain different versions of the same file or database. A slightly different meaning of an archive can be to pack or compress many files together, as in a UNIX tape archive, the very popular tar-file.

In business or government communities, the word archive has all the above connotations, but it has additional meanings. An archive in this sense is a repository of business or government records, and serves the important role of documentary memory for these records. Characteristics of such an archive include preserving the records for a defined term or indefinitely, while maintaining authenticity. In the physical, analogue world, there are specific requirements to ensure this, for example, ISO 15489 (ISO, 2001), but these are not sufficient in the digital world.

[11] OMIM - Online Mendelian Inheritance in Man http://www.ncbi.nlm.nih.gov/omim/

[12] ExPASy - UniProt Knowledgebase: Swiss-Prot and TrEMBL http://www.expasy.ch/sprot/

[13] UniProtKB http://www.uniprot.org/help/uniprotkb

[14] Universität Trier http://dblp.uni-trier.de/xml/

[15] Met Office, UK http://www.metoffice.gov.uk/weather/uk/observations/

Within this paper, we view an archive as a collection of different versions of the same document or database. Therefore, the archiver fits one of the computing-related uses of the term archive, that is, packing many files together, in this case many editions of a related file.

Preservation

Preservation efforts intend to maintain information in a correct and understandable form for current and future use. When we look at preservation in the digital world, the most-cited standard is the Reference Model for an Open Archival Information System (OAIS) (CCSDS, 2002). In the context of the OAIS model, an archive is defined as an organization of people and systems, that has accepted the responsibility to preserve information and make it available. To qualify as an OAIS, an archive must support the OAIS Information Model. It must also obtain *sufficient control* over the data to carry out its functions (this is an important concept, as we shall see). The OAIS model requires the archive to determine for whom it is preserving the information (the *Designated Community*), in order to ensure the information is understandable. It has to follow careful procedures to ensure the authenticity and integrity of the information, using fixity information (e.g., checksums, etc.), and provenance (here meaning records of actions taken in the archive that affect the object).

So what is the OAIS information model? Members of a Designated Community have a *Knowledge Base* that helps them understand some signal as information. In the case of data, that Knowledge Base might suggest use of a certain software package in order to process the data in a way that makes its information content understandable. If this kind of data is unfamiliar to the Designated Community, then somehow it has to work out how to interpret it. For this purpose the archive must supply *Representation Information*. The standard says "Data interpreted using its Representation Information yields Information". In practice for the short to medium term, software will always be the most immediate form of Representation Information. The key to longevity for OAIS lies in the idea that the archive monitors the Designated Community, to determine when its Knowledge Base changes sufficiently that different Representation Information must be provided. This could, for instance, occur when a required software package becomes obsolete, making information encoded by it unreadable. Compared with the traditional archive, these requirements on an archive are onerous, and there are concerns that the high cost of complying with them will result in fewer objects being preserved. In that the archiver is making the set of editions understandable to the user, the archiver itself might be viewed as Representation Information from the practical view described above.

The OAIS model is a useful broad reference, but should not be viewed as a design guide. For many pragmatists, digital preservation means building systems to take in and manage documents and data automatically, providing as much contextual and file format meta-data (a reduced notion of Representation Information) as possible, and loading much of the burden of interpretation on the final user. The Designated Community often cannot be precisely defined, which makes monitoring its Knowledge Base difficult.

Curation

Curation is another loaded term. Our notion of curation in the Factbook is consistent with the approach used in bioinformatics, where the term curated database has been in use for many years (first citation noted is Larsen et al. (1993)). The Wikipedia definition of *biocurator*[16] is "a professional scientist who collects, annotates, and validates information that is disseminated by biological and model organism databases". The role of a biocurator encompasses quality control of primary biological research data intended for publication, extracting and organizing data from original scientific literature, and describing the data with standard annotation protocols and vocabularies that support powerful queries and biological database inter-interoperability. Biocurators communicate with researchers to ensure the accuracy of curated information and to foster data exchanges with research laboratories. In essence, this suggests that curation is the construction of authoritative annotations linking data about significant objects (e.g., genes) with evidence about them in the literature and elsewhere. And this is indeed the sort of thing one sees in genomic databases.

In other parts of science, curation is not so much about linking data to supporting evidence in the literature through the mechanism of constructing annotations on objects, but rather about caring for research data sets (managing them systematically, using community standards, adding descriptions, documentation about how to use them, transferring them to longer-term homes, clarifying conditions of use and provenance information, and also including annotations, etc.), so that they can be used and reused, by the originators or others, now or in the future. This approach incorporates elements of digital preservation, but is not solely defined by long term in the way that digital preservation tends to be.

However, people who maintain curated databases do a mixed job of archiving old versions. The CIA World Factbook is one such example. The CIA presumably keep at least a partial archive of their databases, but it makes no effort to make any past versions available. It is left to an independent group/community (like us) to do it.

Legal Issues

Legal issues can have a major impact. Libraries used to subscribe to business intelligence resources, which were delivered regularly in hard copy form, and which over time built up a significant longitudinal resource. In effect they bought those resources. Now they subscribe to business intelligence databases, for which longitudinal data may be very useful. Subscribe, in this sense, means to lease, for as long as payment continues to be made, under the terms of a licence or contract, which usually excludes keeping additional copies and hence prior versions. Preservation and curation actions always involve making copies and often require making changes, both of which are often restricted by copyright law. In particular, most commercial licences would prohibit the kinds of activities necessary to construct an archive along the lines described in this paper.

[16] Wikipedia http://en.wikipedia.org/w/index.php?title=Biocurator&oldid=

Archiving, Preservation and Curation in Relation to the Database Archiver

How would these notions apply to our database archiver? The primary goal of our archiving effort is to make previous and future versions of the CIA World Factbook available in a unified format under a single query interface. The Factbook is currently published as a set of HTML documents, possibly derived from an underlying database. This database, however, is not directly accessible. Earlier versions of the Factbook are also available, but in different formats. It is not our primary goal to allow exact future recreation of these files or to ensure their long-term readability. Rather, we want to preserve the structure and content that the user sees on paper or on the Web. We therefore maintain the data in an XML text format as this is expected to be readable even over long periods of time. Here is how the other notions from the previous section apply to our archiving effort:

- This paper records how the data were obtained from many different sources. The various editions of the Factbook are in the public domain as US Government productions, so we do have "sufficient control" to carry out preservation actions.
- Our Designated Community is simply the community equipped with current or likely future web browsers and which wishes to read the information.
- The disseminated information from the archiver is the set of web pages produced (in HTML over HTTP) for the user when using the archiver's interface. Both HTML and HTTP are now assured long life due to the extraordinarily wide deployment on the Internet, making them resistant to obsolescence; should they be replaced, future maintainers of the archive would have to provide new Representation Information, in this case a new version of XARCH. The internal structures used in the archive to manage the information, including the different time-encoded variants, are not significant for preservation provided they have potentially good longevity; again XML has this characteristic.
- Because of the range of different ingest sources and formats, the hand manipulation, individually crafted procedures, the unknown history prior to ingest, and the data cleaning required, the authenticity of the information in the archive is uncertain; there is no clear provenance or audit trail. However, since the Factbook is so widely used, should we not preserve it as what people saw and (presumably) believed, and treat it as *raw* data? If we take this view, then "ingest" surely describes what we are doing when we capture successive published versions of the Factbook; and in this activity we do have a clear notion of provenance.
- Since this is a research project, we cannot yet make any institutional commitment to preserving the Factbook. However, we are now running a continuous process that watches for changes in the Factbook and efficiently records them whenever they occur. We believe, therefore, that these techniques contribute substantially to preservation technology for evolving data sets. The degree to which the technology can be applied to other (non-curated) data sets, such as sensor data, that are subject to more volatile change, is still under investigation.

Overall then, we can conclude that the OAIS model and terminology are difficult to apply to XARCH. While XARCH performs a useful preservation function, terms such as "ingest", which have a well-understood meaning in OAIS are difficult to apply to the preservation of curated databases. Moreover, in addition to preserving information for a perilous future, the team has used related technology to obtain information from the threatened past, and made it available for convenient consumption now. It has added significant value to that information.

This notion of adding value for now and the near future is a key feature of curation, one that makes curation different from preservation. The latter relates particularly to resisting information loss over time, and preserving authenticity into the future. In curation, wholesale transformations of the kind accomplished here are accepted if they add sufficient value. In this case, combining past and present (and anticipated future) editions of the Factbook into a single structure qualifies as archiving in the sense we have identified, while supporting this with a simple interface which supports querying and browsing on place, time and information type, represents easily sufficient added value to justify it being regarded as a part of digital curation.

Conclusions

As far as we can tell, the use of the term "curation" in connection with digital artefacts was introduced in the 1990s independently in the worlds of digital preservation and in scientific (mostly biological) databases. In these two contexts the term has substantially different meanings. What we have attempted to show is that elements of archiving can bring added value to curated databases, not only for the purpose of maintaining the scientific record, but also – for a variety of databases – supporting the extraction of temporal information and longitudinal studies. Our experience with the Factbook shows that this is a non-trivial task and indicates that novel techniques are needed, both for the storage and querying of successive versions of a curated database.

At the same time, the accepted terminology of digital preservation is at best confusing when applied to curated databases, and the current practices and standards of preservation need to be substantially revised to deal with this important class of digital artefacts.

References

Buneman, P. (2006). How to cite curated databases and how to make them citable. In *Proceedings of the 18th Conference on Scientific and Statistical Database Management*, 195-203.

Buneman, P., Cheney, J., Tan, W.-C., & Vansummeren, S. (2008). Curated databases. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems,* 1-12, New York, NY, USA. ACM.

Buneman, P., Davidson, S., Fan, W., Hara, C., & Tan, W.-C. (2001). Keys for XML. In *Proc. 10th Int. Conf. on World Wide Web (WWW)*, 201-210.

Buneman, P., Khanna, S., Tajima, K., & Tan, W.-C. (2004). Archiving scientific data. *ACM Trans. Database Syst.*, *29*(1), 2-42.

Cattell, R. G. G., Barry, D. K., Bartels, D., Berler, M., Eastman, J., Gamerman, S., et al. (1997). *The object database standard: ODMG 2.0.* Morgan Kaufmann.

CCSDS (2002). *Reference model for an open archival information system (OAIS).*

Central Intelligence Agency (2009). *The world factbook.* Washington DC: Central Intelligence Agency. Retrieved August 3, 2009, from https://www.cia.gov/library/publications/the-world-factbook/

Galperin, M. Y. (2008). The Molecular Biology Database Collection: 2008 update. *Nucl. Acids Res.*, *36* (Database Issue), D2-D4.

ISO (2001). *ISO 15489 Information and documentation — Records management*

Koltsidas, I., Müller, H., & S. Viglas. (2008). Sorting hierarchical data in external memory for archiving. In *VLDB '08: Proceedings of the VLDB Endowment*, 1205-1216.

Larsen, N., Olsen, G. J., Maidak, B. L., McCaughey, M. J., Overbeek, R., Macke, T. J., et al. (1993). The ribosomal database project. *Nucl. Acids Res., 21*(13), 3021-3023.

Müller, H., Buneman, P., & Koltsidas, I. (2008). XARCH: Archiving scientific and reference data. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1295-1298.

Schneider, M., Bairoch, A., Wu, C. H., & Apweiler, R. (2005). Plant protein annotation in the UniProt Knowledgebase. *Plant Physiol.*, *138*(1), 59-66.