

Title	3.空値とその始末 : 半構造の設計論(1 ジエネティック・アーカイヴ・エンジンの方法論,ジエネティック・アーカイヴ・エンジン : デジタルの森で踊る土方巽)
Sub Title	
Author	遠山, 元道(Toyama, Motomichi)
Publisher	
Publication year	2000
Jtitle	Booklet Vol.6, (2000. ) ,p.38- 45
JaLC DOI	
Abstract	
Notes	
Genre	Journal Article
URL	<a href="https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AA11893297-00000006-04394202">https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AA11893297-00000006-04394202</a>

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the Keio Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

### 3. 空値とその始末

——半構造の設計論——

遠山元道

半熟卵って、つまるところ生卵でなくて、固ゆで卵でもない、卵のスペクトルです。

半構造って、つまるところ無構造でなくて、強構造でもない、構造のスペクトルなんです。

「万人においしい半熟卵」がないように、半構造のアドホックで規範性の薄い所が工学的には厄介ですね。

#### はじめに

半構造データモデルは柔軟で、イレギュラーな情報の記述に適している<sup>★1</sup>。従来の強構造のデータモデルではイレギュラーな情報に遭遇するとしばしば立ち往生をしてしまうが、半構造ではとにかく記録ができる。しかしながら記述の柔軟性は諸刃の剣で、下手に振り回せばわが身を傷つけてしまう。ここでは仮想アーカイヴの記録を例として、頻繁に生じ、また重要度の高いイレギュラリティである空値（およびその裏返しの多値）について考える。

#### 例示

いささか簡略に過ぎるが、絵画の書誌情報を例にとってみよう。

##### ・無構造（文章）<sup>★2</sup>

1901年から1904年に描かれた作品のうち、どうしても紹介されるべきは、以下のものである。紙の上のパステル、優しさが溢れた「病んだ子供と母親」——バルセロナ・ピカソ美術館蔵——、クリーブランド美術館所蔵のキャンヴァスに描かれた油彩で、「青の時代」の作品のうち最も有名なもの一つ「人生」。

・強構造（関係データモデル）

id	作 者	タ イ ル	年	画 材	サ イ ズ	所 �藏
172	パブロ・ピカソ	病んだ子供と母親	?	紙にパステルとコンテ	46×40	バルセロナ・ピカソ美術館
188	パブロ・ピカソ	人生	1903	カンヴァスに油彩	?	クリーブランド美術館
256	パブロ・ピカソ	侍女たち	1957	カンヴァスに油彩	194×260	バルセロナ・ピカソ美術館

・半構造（XML）

```
<絵画 id=172>
  <作者>パブロ・ピカソ</作者>
  <タイトル>病んだ子供と母親</タイトル>
  <画材>紙</画材><画材>パステル</画材><画材>コンテ</画材>
  <サイズ>46×40</サイズ>
  <所蔵>バルセロナ・ピカソ美術館</所蔵>
</絵画>
<絵画 id=188>
  <作者>パブロ・ピカソ</作者>
  <タイトル>人生</タイトル>
  <年>1903</年>
  <画材>カンヴァス</画材> <画材>油彩</画材>
  <所蔵>クリーブランド美術館</所蔵>
</絵画>
  . . .
```

### 空 値

人のアクセスし得る情報には限りがあり、歯痒い思いをしながらも分からることは分からないで済ます他はない。強構造の関係データモデルでは、表の構造（スキーマ）が固定であり、記録のインスタンス（タプル）はスキーマに従う。無構造の文章や、半構造の記述が可能な XML (eXtensible Markup Language) では分かっていることだけを記述すれば良いが、表では情報の欠損部分は空欄にしておくことになる。

空値についてはさまざまな解釈が可能である。人がこれを見て誤った解釈をするかもしれない。さらに微妙で深刻な問題として、空値を含むデータベースを対象とした質問処理、統計処理の結果、欠損が形を変えて虚偽の事実に化けてしまうこともあり得る。これらの問題点について、C. J. Date によって最初に詳しい検討が加えられた<sup>★3</sup>。

### 解決しない空値

空値を必要とするのは未知、すなわち解決が予定される空値のほかに、

値がそもそも存在しない場合がある。ある絵画が焼失してしまったとすると、収蔵場所については値が永久に得られない。空値には、大別すると未知 (unknown) の空値と、不在 (non-existent) の空値の2種類がある。ここでは表に記入する際に、前者を疑問符 (?)、後者をダッシュ (—) で表することにする。

不在の空値をさらに観察すると、上記のような例外的なケースで発生するものの他に、不適 (inapplicable) の空値、すなわち特定のクラスの対象にある属性の値が存在し得ないために発生する空値というサブカテゴリに気づく。彫刻のサイズは高さ、幅、奥行を記述するが、絵画に奥行は不適な属性で、値が存在しない。

### 問題の解決、先送り、消滅

半構造データモデルでは情報の欠損に、「記述しない」という形で対処できる。しかし、記述しないことで問題が解決してしまうことは稀で、多くの場合問題を先送りしているに過ぎない。表に空欄があれば、見るものに情報欠損の事実が伝わる。記述しないこと、によって先送りしたはずの問題が視界から消え去り、未解決のまま消滅してしまう恐れもある。このことに配慮し、未知の空値についてもタグを設け、“To be supplied” のような特殊な値を仮に記入することもしばしば行われる。

この方法は半構造ばかりでなく、強構造の関係データベースでもデフォルト値という概念で標準言語SQLに採用されている。デフォルト値は覚え書きとして役立つが、きちんと意識しないと質問処理において「統計のウソ」問題に荷担する危険がある。

### 多 値

強構造の関係データモデル<sup>★4</sup>では、E. F. Coddの規定した第一正規形であるための制約から、各々の属性値はアトミックであることが要求される。画材の欄に、「紙にパステルとコンテ」のように書いてあるのは、これが一つの文字列のため、シンタックス・チェックを逃れ得たにすぎない。明らかにここには三つの情報を含み、セマンティクスはアトミックでない。従って上記の表は関係データモデルにおいては第一正規形条件を満たさない誤った設計である（関係データモデルにおける多値の正しい扱いについては後述する）。XMLによる表現例では画材タグを繰り返し使用することによって、多値属性を自然に表現している。

### 部分情報としての多値

空値にさまざまな解釈があるように、多値の解釈も一意的ではない。連言の多値は複数の値がどれも真である記述で、制作年が1901年、1902年という多値ならば、これは2年に渡って制作したことを表す。選言の多値は複数の値のいずれか一つだけが真という解釈である。上記の例で、制作

年に関して文章から得る以外に情報がなければ、「病んだ子供と母親」の制作年は1901、1902、1903、1904と多値にしなければならない。これはアトミックな整数でないので、シンタックス・チェックにかかり、関係データベースには記録できない。実際の値1903年が不明ならば、1901-1904年というレンジが有用な情報となることもある。このように、ある事柄について十分ではないが何らかの情報が得られる場合、これを部分情報(partial information)または不完全情報(incomplete information)とよぶ。部分情報は、多値で明示的に与えられる他、次項で述べるように論理的な推論で発生するもの、両親の血液型から子供の血液型を生理学的な法則に基づいて推論されるものなど、様々な原因から生じる。

### 部分情報の形式モデル

部分情報については、W. Lipski Jr. によって1979年に初めて不完全情報代数による形式化がされた<sup>★5</sup>。Lipskiは、部分情報を属性の持ちうる値の集合としてモデルを構築した。既知の値、すなわち完全情報は、要素をちょうど1つもつ集合で表す。空値は空集合で表すように思えるが、逆にその属性の定義域を規定する集合全体をもって表す。ある人の血液型が不明ならば、それは空集合  $\emptyset$  ではなく {A, B, AB, O} という集合で表す。したがって、集合の包含関係が  $S \subset T$  ならば、S が T より情報量が多いことになる。集合の包含関係は代表的な半順序関係であり、また任意の2要素について結びと交わりが定義できることから全体として完全束を構成する。最大元が空値、単元集合が完全情報を表すことは明らかである。集合から元を減らして小さくすることは情報量を増すことになり、これを属性値の完全化(completion)と呼んでいる。

### 関数従属性から推論される部分情報

関数従属性 (Functional Dependency) は、関係データモデルの設計理論の中心的な概念で、これに基づいて第2正規形、第3正規形、Boyce-Codd 正規形などが定義される。関係 R(A, B, C) において、関数従属性 A → B が成立するのは以下の場合であり、またその場合に限る。

$$\forall s, t \in R \ s.A = t.A \supset s.B = t.B$$

たとえば属性 A が学生の学籍番号で属性 B が氏名とすると、任意の2学生 s と t について、s と t の学籍番号が一致するなら、それらの氏名も同一でなければならないということを表している。このことから、関係 R において関数従属性 A → B が定義されていると、空値について以下の2通りの推論が可能である。

前方推論：タプル t1 (a1, b1, c1) とタプル t2 (a2, ?, c2) が存在し、かつ a1=a2 ならば、t2 の属性 B (未知) の真の値は b1 である。

後方推論：タプル t1 (a1, b1, c1) とタプル t2 (?, b2, c2) が存在し、かつ b1 ≠ b2 のとき、t2 の属性 A の値は「a1 ではない」とい

う部分情報をもつ空値となる。

### 名前つき空値

関数従属A→Bの成立するデータベース中に、属性Bの値が相異なる複数の記録で属性Aの値が空値だったとする。上記の後方推論を一般化することで、これらの空値の実現値（既知になったときに与えられる真の値）が互いに異なることが導かれる。これらが異なることを明示するためには、空値の記号を複数使用して、( ? 1, b1), ( ? 2, b2), …としたうえで、 ? 1, ? 2, …が互いに不等であることを述語論理式で記述することができる。このような空値を名前つき空値（named null）と呼んでいる。名前つき空値は、実現値の代理としてデータベースに対する統計的質問に正しく振舞うことが利点として挙げられるが、一方で空値を空値として扱いたい質問においては処理をきわめて煩雑にする欠点がある。

### 部分情報と質問処理

Lipskiのモデルでは、データベースが部分情報をもつとき、これに対しても与えられる問い合わせQに対する解答として、極大解（maximal answer）と極小解（minimal answer）を定義する。極大解 $\| Q \| ^*$ は質問Qに最も楽観的に答える解で、極小解 $\| Q \| _*$ は悲観的、もしくは確実な答えである。「血液型="A"の人の氏名を求める」質問に対し、前者では部分情報{A, O}を持つデータを解に含めるが、後者では含めない。一階の述語論理式で与えられる任意の質問に対し、 $\| Q_1 \wedge Q_2 \| ^* = \| Q_1 \| ^* \cap \| Q_2 \| ^*$ のような書き換え規則によって質問を簡略化し、解を計算する体系を与えている。

Q1が血液型="A"、Q2が血液型="O"で、質問Qが $Q_1 \vee Q_2$ の場合に極小解 $\| Q \| _*$ を求めようとすると、データ{A, O}は確実にQを満たすので答えに含めなければならない。しかし、これらを構成する部分質問Q1、Q2のいずれの極小解にもこれが含まれないことからも分かるように、質問分解／解答合成のプロセスは見かけほど単純ではない。

### 統計的質問と空値

関係データベースの質問言語SQLなどにも、データベースの統計的性質を質問するためのプリミティブとして、count(), average()などの集約関数が用意されている。データベースが空値や多値などの部分情報を含む場合に、このような質問に対する解答はしばしばウソを含んでしまう。

特定の作者の絵画の点数を調べるときに、select count(id) from R where 作者="XX"と質問したときに、たとえばこのうち3件のid属性に空値があり、これらが同一視されると正しい点数より2点低い値が返されてしまう。前述の名前つき空値はこの問題を解決するのに役立つ。

つぎに、絵画の記録が10件あり、そのうち3点については購入価格が記

録されていて他の7点については空値のとき、購入価格の平均を分母を10として計算すると不当に低い値が返されてしまう。これを避けるためには購入価格が空値でないものを選択してから集計すればよいが、この場合には名前つき空値は処理を複雑にしてしまう。

### 3 値論理

普通、論理で扱う真理値は真と偽の2通りで、これに基づいて論理演算の真理値表は $2 \times 2$ のマトリックスで与えられる。しかしながら、データベースが空値を含む場合に、「収蔵="プラド美術館"」のような単純な述語の真理値が真とも偽とも定まらない。これについて、空値を対象とした述語の真理値を不明とし、真、不明、偽の3通りの真理値を用いる3値論理が関係データベースの標準言語SQLで採用されている。真理値表は $3 \times 3$ のマトリックスとなる。連言と選言の真理値表において、

真 $\wedge$ 不明=不明、偽 $\wedge$ 不明=偽、真 $\vee$ 不明=真、偽 $\vee$ 不明=不明  
と定義される。述語論理の限量子については、 $\forall$ は $\wedge$ 、 $\exists$ は $\vee$ に準じて定義される。

### DTDと記述の規範性

XMLでタグを用いて書けばどのような構造も自由に書けてしまう。DTD(Document Type Definition)は、XML文書に現れるタグに関する構文規則を与えるもので、見方によればこれは関係データベースなどの強構造データモデルのスキーマに相当する。ただし、XML文書がDTDに従うことには強制されない。XML文書を構文解析し、DTDに準拠しているかどうかをテストすることはできる。このため、DTDは文書記述の規範としての役割を担うが、例えば<タイトル>というタグを用いるべきところに<題名>と書いてしまっても、明示的にチェックを行うまでは発見されない。どのようなタグを使用すればよいかが分からぬときにこれを参照することで、DTDをデータガイドとして用いることもできる。

### コントロールされた先送り

関係データモデルでは、前述のように属性値はアトミックに限り、多値を含む複合的な表現を許していない。では多値をどのように扱うのだろうか？詳細は省略するが、スキーマに含まれる関数従属性の構造に着目し、表を分解することによって、多値などを除いたBoyce-Codd正規形(BCNF)のスキーマへ必ず変換可能であることが証明されている。冒頭の絵画情報の例で、画材属性は多値となる。元のスキーマR(id, 作者, タイトル, 年, 画材, サイズ, 所蔵)から多値属性である画材を分離し、R1(id, 作者, タイトル, 年, サイズ, 所蔵)とR2(id, 画材)に分解することによって、これをBCNFにすることができる（図1）。R1とR2から関係代数の結合(join)演算によって、元のRを復元することができる。

R1

R2

R3

図1 絵画情報のBoyce-Codd正規形スキーマ

さて、空値は多値と一見逆の概念のようだが、多値とは値をいくつ持つてもよいことを表し、これはゼロ個のケースを含む。したがって、ある属性がしばしば空値を持つ場合にこれを多値属性と同様に扱い、BCNFへの正規形分解を行うことがある。この際、空値の側の関係には対応するタプルが無くなる。絵画の記録で、一部にだけ解説が与えられていたとすると、多くの絵画について解説属性は空値となる。これを別の表に分けて、R3(id, 解説)とするのは正しい設計とされている。R3は解説の数だけタプルを持つ。R3をR1、R2と結合すると、解説の無い絵画のタプルは結果から排除されてしまう。相手のタプルが存在しないときに、これを空値で補ってすべての組合せを答えに含める特殊な結合演算が定義され、外結合と呼ばれている。図1の例では、R1とR3の通常の結合ではid=256のタプルだけが解に含まれるが、外結合では結果は3タプルとなり、id=172とid=188の解説属性は空値となる。

対象を図1のR1,R2,R3のように分解してモデル化し、これを内結合（通常の結合）するか、外結合するかを検索の実行時に選択することは、空値、多値の諸問題をデータ記述時に解決するのでなく、検索時に情報利用者に選択に委ねるコントロールされた先送りの方法論といえる。

### ユニフィケーション

以下の2件のレコードが存在したら、一般にこれらをどう扱うだろうか？

(id=75, 作者="ディエゴ・ヴェラスケス", タイトル="侍女たち",

年=1656, サイズ=?, 画材=? , 所蔵="プラド美術館")  
(id=? , 作者="ディエゴ・ヴェラスケス", タイトル="侍女たち",  
年=1656, サイズ="318×276", 画材="カンヴァスに油彩", 所蔵="  
プラド美術館")

両者にはそれぞれ未知の空値があり、無矛盾である。ここで無矛盾とは、対応する既知の値同士が一致していることを指す。この例ではすべての未知に既知値が対応するので、さらに補完関係にあると言える。無矛盾の関係にある記録をまとめて一つの記録で置きかえることは、Horn 節に基づく自動推論機構で用いられるユニフィケーション (unification) と等しく、データベース総体としての論理的な意味を変えることはない。

巨大なデータベースでは、追加情報を既存の情報と照合してユニフィケーションを自動的に行う機構も欲しくなる。しかしながら、空値が多用される場合には自動的なユニフィケーションは明らかに危険を伴う。"パブロ・ピカソ／侍女たち／1957／バルセロナ・ピカソ美術館"という共通項目をもつ記録が数十もあれば、自動ユニフィケーションによっていくつかが単一化され、必要な情報が失われても不思議でない。

アーカイヴのように管理された記録ではなく、インターネットに散在するWebサイトの情報を統合的に利用するようなシステムでは自動ユニフィケーションは非常に有用である。この場合には無矛盾は制約としてむしろ強すぎて、表記のゆれ (ヴェラスケス、ベラスケス、Velasques、ディエゴ・ヴェラスケス……) を同一視することなどが必要となる。

### 註

- ☆1—— S. Abiteboul, "Querying Semi-Structured Data," *Proc. Int. Conf. on Database Theories*, 1996.
- ☆2—— G. C. Cravel, "ピカソ：バルセロナ・ピカソ美術館（日本語版）," 31 ページより。
- ☆3—— C. J. Date, "Null Values in Database Management," *Proc. 2nd British National Conference on Databases*, 1982.
- ☆4—— E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of Association for Computing Machinery*, Vol. 13, No. 6, 1970.
- ☆5—— W. Lipski, Jr., "On Semantic Issues Connected with Incomplete Information Databases," *ACM Transactions on Database Systems*, Vol. 4, No. 3, 1979.

(とおやま もとみち・慶應義塾大学専任講師／情報工学)