

人文資料の符号化向け分析・評価基準の枠組み

大矢 一志
鶴見大学

本稿では、人文資料の符号化事業で採用できる分析・評価基準の枠組みを整理する。人文情報学の対象となる電子化資料 (a) 生来デジタルオブジェクトと (b) 電子化オブジェクトのうち、後者は (x) テキストと (y) 非テキストが対象となり、(x) で作られるオブジェクトには (i) テキスト版、(j) ファクシミリ版、(k) アプリケーション版がある。このうち (i) については、(1) メタデータ/アノテーション、(2) データ単位、(3) データ構造 (スキーム) が符号化の分析・評価基準の対象となり、その分析・評価の手法には、システム開発の要求分析の手法が適応可能である。これを可能にする背景には、人文情報学とメタデータの歴史から得られるアノテーションの位置づけがある。また、(a)(j)(k) についてもシステム開発の手法を採用することが可能である。(y) はモデル研究における手法を採用することが可能である。(a) の派生系として、(a)(b) を元に二次資料として作られるオブジェクト (c) があり、これにはシステム開発の手法、モデル研究の方法、人文学における二次資料の評価法を採用することができる。

A Story of Analysis/Evaluation Criteria for Digitization of Humanities Resources

OHYA Kazushi
Tsurumi University

This paper shows a framework of analysis/evaluation criteria to be used in digitization or making digital projects especially for digital humanities(DH) studies. The objects as targets of DH are (a)born-digital objects, (b)digitized objects, and (c)secondary objects made from (a) or (b). The (b) is made from (x)text content or (y)non-text content such as maps, pictures, sound and so on. The object (x) can be categorized into (i)text version, (j)facsimile version, and (k)application version. As a way of analysis/evaluation for making digital objects, a way used in system science, especially requirements analysis, and a way in a model study can be adopted in the case of (a),(j), and (k), and in the case of (y) and (k) respectively. In the case of (c) we can adopt a way either in system science, a model study, or a traditional humanities studies. The (i) has three targets of analysis/evaluation, i.e. an annotation, a unit, and a structure or scheme that can be aliased/evaluated in a way of a model study.

1. はじめに

本稿は、人文資料の電子化に関わる事業で符号化の設計を検討する際に利用できる、分析・評価基準の枠組みを提示することを目的とするものである。特に分析領域のひとつとしてあるメタデータ/アノテーションを詳しく展開している。資料を電子化する際に採られる分析・評価基準は、いわゆる口伝で伝えられる、または経験から総合的な気付きで得られる知識とされるようで、人文資料の電子化を扱う書籍でも、広くまとめられたものを知らない。その背景には、取り扱う範囲の広さと、技術や研究トレンドが刻々と変化してきたこともあるのだろう。本稿にまとめた内容は、工学寄りであれ、計算機科学寄りであれ、人文学寄りであれ、概論の知識として共有できるものと考え。それぞれの領域では、当然すぎる内容であることも、書籍化されてこなかった理由としてあるのかもしれない。いずれにせよ、新しく人文情報学(digital humanities)に参加する研究者が、気づきの経験に一定の年月を必要とすることから習得が遅れる傾向にあった知識である。そこで、本稿をかりてこれらをまとめておくことにした。はじめに、資料が電子化されてきた歴史と、メタデータの歴史を振り返り、そこから資料の符号化ではアノテーションが評価・分析の中心的な対象となることを確認する。次に、評価・分析対象としてある電子オブジェクトを整理し、符号化の検討基準として一

般に挙げられている項目を整理する。次に、Ontological Semantics と呼べる古い手法による新しい研究を確認して、最後に、テキスト版と分類されるオブジェクトを取り上げ、評価・分析基準を確認する。

2. 符号化されてきた資料

人文情報学の発展史は、資料の電子化・符号化の歴史である。極論かもしれないが、主立った成果は電子資料しかないとするのが潔い評価であろう。人文情報学は、一般の科学研究のように、何が発見されたか、何に使われたのかで研究史を辿ることは難しい。但し、符号化された資料から、見るべき発見や成果がなかったわけではない。そのような成果を生んだ資料の代表例としては、言語資料(corpus)と電子学術資料(digital scholarly edition)を挙げることができる。この2つは、発展史上、人文情報学の性格を作り上げてきたといえる。

コーパスは、辞書作りの基礎データとしての役割に加えて、従来までの辞書作りでは表に出てこない元資料そのものをユーザが扱う機能も期待され、作られてきた。その最新の様子は、COCA(the Corpus Of Contemporary American english)に見ることができる。但し、コーパスが辞書の元資料としての役割を超える機能を持つことを示したものは、現在でもBNC(British National Corpus)において他にない。BNCではコーパスの引用元のメタデータとして、従来の書物向けの書誌データにはない情

報, 例えば, 著者の年齢, 性別, 社会階層, 文体などを記録し, その情報から検索用語の使用状況を数値で知ることが出来る。このようなコーパスは, 自然言語処理研究のように, 規範文法を求める, すなわち, より多くの人が納得する (使用している) ものを決定するためだけに使われるのではなく, 語の利用状況を知ることができる, いわゆる記述的言語研究を支える資料として使用することができる。BNC は, メタデータにアノテーション, すなわち著者以外の作者による記録を中心的な役割として持たせたことで, 自然言語処理研究とは異なる独立した研究領域を発展させた重要な成果である。

もうひとつの代表的な成果は, 電子学術資料と呼ばれるもので, 研究資料として作り出されたものである。人文学では, 主として書物を対象に研究がされているが, 研究のために書物の書物いわゆる二次資料が作られ, その二次資料を作るメタ研究活動自体も研究活動と見なされてきた歴史がある。従って, このような電子学術資料を作成すること自体も研究活動となり, 人文情報学では中心的な活動になっている。この二次資料は, 書物を直接研究対象とするのでは手間がかかる, 書物間の関連性という, ひとつ抽象度の高いテーマに関心を持つ研究の為に作られた, 複数の資料の関係を整理した資料である。従って, テキストの持つ線状性というメディアの制約や, 印刷物という物理的制約は, 障害となることが多く, その克服のために様々な工夫が施されてきた。電子学術資料とは, 「書物間の関連性の表示」を, ハイパーリンクを使い, 閲覧性を理想までに高めた研究資料を電子版として実現したものである。このような二次資料の作成では, 資料作成そのものが研究目的となるため, そこから何が発見されるのかということ, 第一の評価点にはならない¹。ところが, 不幸なことに, 電子学術資料は, 長い間, 人文情報学以外のコミュニティからは実質的な良い評価を得ることができなかった。具体的には, a) 電子学術資料は実質の寿命が, 利用環境の変化から数年程度しか持たない資料であったこと, b) それが原因となり二次資料の実質的評価²を決めてきた図書館から低い評価が与えられ, c) 結局, 電子学術資料の作成が難しくなったことがある。但し, 近年は, web 上での利用を想定した電子学術資料は利用環境の変化に強いと想定され, 再び資料作成が盛んになってきている [1, 21]。更に現在では, 電子化された資料を元に作られる電子学術資料が, 人文情報学で作られるオブジェクトの主流になろうとしている。

これらコーパスと電子学術資料の成功の歴史から, 人文資料の電子化で採られる分析・評価基準の枠組みを検討する指針を得ることができる。これらの詳細は, 本稿では分析対象を整理した後に示されるが, 先にまとめると, a) メタデータはアノテーションであるべきこと, b) 電子学術資料は存在そのものに価値が置かれてきたことから, 独善的ともいえなくもないが, 少なくとも, 極めて特定のコミュニティから評価を得られるだけで成立できることが指針となる。

3. メタデータの歴史

メタデータの定義を, 電子オブジェクト一般の内容や特性に関する情報, すなわち “data of data” として, 書籍に対する書誌情報に相当する情報とすれば, その性格や役割を理解するために, いわゆる書誌情報の歴史を確認することは有益である。結論を先に言えば, 現在のメタデータには, 発展史上, アノテーションの機能が求められている。

ローマの時代, 図書館で収蔵されていた巻子は, 書かれている言語であるギリシャ語とラテン語により部屋が分かれていた。ここでは, 部屋割りが書物を分類する役割を果たしていたといえる。現在の図書館でもそうであるように, 書物の分類は物理的な配置で示すことができる。また, 巻子を収蔵する場合, 巻子を紐解かないと内容が解らないようでは不便であることから, そのタイトルを記した「タグ」が巻子に付されていた。このタグは, 書物に代わる役目を持ち, さらに (1) 書物の名前を示す役割という “data of data” の役割を持つことになる。このタグを巻子から切り離し, カードに形を変えて一カ所に集めると, いわゆる目録カードが生まれる³。書物から切り離された目録カードには, タグが担う書物に代わる機能が強化され, (2) 書物の存在を証明する機能と, (3) 書物の場所情報を示す役割が加わった。

(1) 書物を特定する識別子としての役割を果たすことが期待される書物名や著者名は, その情報のみで識別子としての役割が果たされる訳ではない。名前相当の情報と物理的存在との一致確認, すなわち識別子解決は, 人が担うことで果たされる。これにより, 書名・著者名などの名前は, 識別子としての役割は解決する人の能力に頼るという意味で不完全な機能しか持たないことになるが, 一方で, この不完全さは名前を曖昧な識別子として運用することを可能にしてきた。この不完全さのおかげで, 写本における物理的な存在と作品としての形而上の存在が混在する識別子を利用することが可能となり, 現在に至る書誌学/文献学の調査を可能にしてきたといえる。現在の書籍においては, 複数の特定項目⁴が識別子を構成する要素として規格化され, 写本のそれよりは識別子解決が容易になっているといえるが, 目録カードにある識別子は, 人の認定が伴うことから識別子としての機能はやはり不完全である。

(2) 巻子や書物から切り離され, 独立したタグや目録カードは, 書物の存在そのものの代理として存在し, カードの存在が書物の存在証明となる。当該書物が存在しなくなれば, その代理としてのカードもなくなり, 当該書物は存在しないことが示される。書物の存在と同等であることは, 目録カードそのものの価値を高めることになり, カードそのものの管理が重要視されることになる。

(3) 目録カードには, 人による識別子解決を補助するために, 書物が物理的に存在する場所の情報が記されている。ちなみに, 人による場所解決には, その人が示された場所にたどり着くことが前提とされていることから,

¹ もちろん, 使える・使えないという二次的な評価は可能である。

² 購入するということ。

³ 書物の分類, 目録, 目録カードの出自についての正確な歴史は図書館学の専門書を参照して頂きたい。

⁴ 題名, 著者名, 出版年, 版番, 出版社など。

その人の能力や権利により、場所解決の実効性が制限される。また、当該場所にたどり着いた場合においても、その人の能力や、書物の管理のされ方により、識別子解決の実効性も制限されてくる⁵。

このような機能を持った目録カードが書誌データとして図書館システムに取り込まれると、(1) 識別子は検索機能と共に提供され、(2) 存在証明の役割は同じく、データの存在が書物の存在を示し、(3) 場所解決を補助する情報提供も同じく担うことになるが、新たに(4) 内容に関する情報が注記として加えられるようになった。この内容に関する注記は、分類とは別に、書物の内容に関する情報として作者以外の人物による記述が情報として加えられたものである。ここに記述される情報には、要旨、宣伝文句、目次の他、著作権情報などがあり、書物本文の意味内容に関する項目というよりは、想定を超える書誌情報を記述する項目としての役割を担っているといえる。

書誌データに新たに加えられた内容に関する情報は、書物の内容に関するとはいえず、その作者は著者以外の第三者である。内容に関する情報を目的とするのであれば、書物の内容そのものが望ましく、ここに書物が電子化される要求が生まれてくる。同時に、著者以外の第三者の記述が果たすべき役割は、独自の視点による要旨や、正・負いずれの批評を記録することよりも、人文情報学で BNC が示したようにアノテーションの記録の方が望ましい⁶⁷。また、図書館システムよりも規模の大きいインターネットの歴史からも、内容に関する情報として求められるものは、書誌データに加えられた内容に関する注記ではないことが分かる。

インターネットでは、オブジェクトの識別子としての URL、書物の識別子に相当する名前解決としての URN、内容に関する情報としての URC が検討され、同時に、これら識別子とオブジェクトとの厳密な一致処理(解決)が検討されてきた。書物が電子化されオブジェクトとなると、(1) や (3) で必要となる各種の解決はシステムが担い、システム内の名称や場所の識別子は、オブジェクト実体と明示的に関係づけられる必要がある。識別子はユニークであることが求められ、解決処理を人が担っていたときのような、曖昧さ・不完全さは許されない。インターネットの歴史の結果としては、名称と場所識別子が一体化した URL⁸のみが解決処理対象とされ、URN と URC はそれから外された。その背景には、1) URN や URC の定義が難航していた、2) ネットワーク全体に影響する特別な仕組みを新たに導入する必要性が認められなかった、3) Google が示したキーワードを問い合わせとする解決方法がそれなりの効率で同定(一致処理)が実現されて

いた⁹、等がある。結果として、インターネット上では、メタデータに識別子の役割は期待されていない¹⁰。また、(2) 存在証明は、当該オブジェクトが担うことが可能で、この役割をメタデータに求める必要も無くなった¹¹。インターネットの発展史上、識別子としての役割が求められなくなったメタデータには、書物のオブジェクト化と、メタデータ内にアノテーションの記述が求められることになった¹²。

3.1 アノテーション

改めて確認をすると、アノテーションとは、当該作者または制作者以外の人物による記述(注釈)で、主に、当該符号化する人またはプロジェクトにおいて作成されるデータである。人文情報学のコーパス作成では、コーパスを利用する場面を想定して、利用者が観察したい情報を、アノテーションとしてメタデータとして記述した。すなわち、何を符号化するのかという分析・評価基準では、何を観察したいのか、すなわち利用者の要求が分析の対象となり、アノテーション一般の評価基準になっている。これが、一般オブジェクトの作成においても同じであるとすれば、人文情報学で作られる各種の電子資料¹³においても、同様に分析・評価の対象と手法を導入することになる¹⁴。

4. 対象と手法の分類

分析・評価の手法を検討するために、その対象となる人文情報学が対象としてきたオブジェクトを整理すると、(a) 制作の時点から電子物である born-digital(生来デジタル)objects と、(b) 物理的な媒体から電子化された digitized objects の2種類のうち、(b) digitized objects では、電子化の対象物として (x) テキストと、(y) 非テキストの2種類があり、さらに (x) テキストを対象としたものは、電子化されるオブジェクトの種類により (i) テキスト版、(j) ファクシミリ版、(k) バイナリ/ソフトウェア版の3種類に分類される。さらに、現在では、(a) born-digital objects または (b) digitized objects を元に、二次的な資料として作られたオブジェクト (c) も対象とされている。まとめると、分析・評価対象は、(a)(i)(j)(k)(y)(c) の6

⁹ 面白いことに、人が解決処理を担っている。

¹⁰ メタデータにある識別子に完全な解決が期待されなくなったということで、メタデータそのものを含むオブジェクトの識別子が不要になった訳ではない。

¹¹ 図書館学で論議される、レポジトリにおいて、収蔵されていない実オブジェクトの代わりにメタデータを使用するケースや、場所解決が不可能なもの、例えば、アクセスの制限により直接の解決ができない場合を想定した存在代理の役割をメタデータに求めるケース、さらに、オブジェクトそのものの保存は諦めメタデータのみで保存を考えるケースにおいては、メタデータに存在証明の機能を求めることがある。

¹² 但し、後述するように、ここでは単にオブジェクトとしているものには、digitized objects と born-digital objects があり、前者の場合、元資料としてある書物のメタデータでは、依然として (1)(2)(3) の機能が期待されることから、現物とオブジェクトの関連性に関心を持つ場合には、ここに新たに、計算機を記録に使うことで新たに生まれる、いわゆるドキュメンテーションの課題が生まれることになる [19]。

¹³ この分類の詳細は次章にある。

¹⁴ ちなみに、TEI P5 では、著者による注釈と著者以外による注釈を独立した要素として分けていない。TEI P5 では、要素 note が属性でそれを区別するアプローチを採っている。あまりよいデザインではない。

⁵ 例えば、ロシア語を読めない人は、ロシア語で書かれたタイトルを探すことは出来ない。また、書物が棚の後ろに落ちていたりする管理体制下では、人が識別子解決することに相当の労力が必要となる。

⁶ 批評とアノテーションは異なる。理由は、後述するように、アノテーションは要求分析から求められるが、批評はそうではない。

⁷ 例えば、これにより特定キーワードを多く含む文書の作者の傾向が分析できる。また、アクセス要求元が記録されていれば、文書の利用傾向が同時にわかり、利用者が所属するコミュニティで引用するに相応しい文書であるかどうかを、検索結果と同時に知ることが出来る。

⁸ プロトコルなどのアクセスに関する情報のパッケージ。

種類となる (図 1).

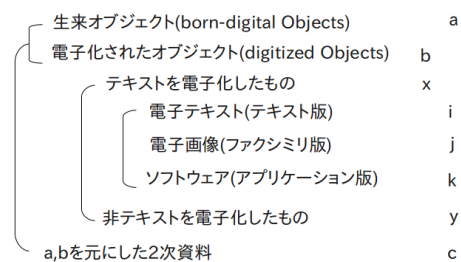


図 1 分析・評価の対象

これらのオブジェクトのうち、(a) の生来デジタルオブジェクトの製作には、電子書籍でもアプリケーションでも、ソフトウェアを使い作られる以上、システム開発の工程が採られる。従って、その製作物の分析・評価は、利用者の要求を分析する手法そのものとなる。また、(j) ファクシミリ版も (k) バイナリ版もソフトウェアで作られ、また (k) ソフトウェアはそのものであるから、これらの分析・評価もシステム開発で採られる要求分析の手法を使うことができる。

(y) 非テキストを対象として電子化されたオブジェクトは、モデル化の手法を採ることから、モデルの分析・評価基準をそのまま採用することができる^{15,16}。

(c) は、従来の人文情報学では扱われず、これからの主流となる対象であることから、後述する補足にもあるように、本稿で扱うことは難しい。現状では、モデル研究における分析・評価基準と、従来からの二次資料そのものの存在価値の判定基準が採用できると考えられる (cf.6 章)。

4.1 要求分析

情報システム、すなわち計算機を人の組織的な活動に導入した仕組みを、作成する、いわゆる (情報) システム開発では、人、組織、計算機の三体¹⁷を定義・固定して、ソフトウェア上に実装する必要があることから、システムの設計は一意に決まらない。従って、理想解を求めるプロセスとなる開発手法は事業の成果に大きな影響を与えることから、これまでに多くの開発手法が検討されてきた [20]。但し、いずれのシステムの分析・評価の手法でも、要求分析が基本となることは共通している。システム開発では、利用者の要求やその環境を分析し (要求分析・要求定義)、まとめられ (仕様書・要求定義書)、それを元に開発 (設計、実装) されて、完成したシステムは、分析された要求がそのまま評価項目となり、評価される。

つまり、製作物の分析・評価対象は、利用者の要求そのものになっている。システム開発では、利用者の要求を軸に、設計 (アーキテクチャ) を求める手法が採られている。

資料を符号化する場合、時にその利用者である人文学者が、何を求めるのかを明確にできず、要求がまとめられないことがあるが、拙著 [13] でまとめたように、これは参加者の資質を問う人文情報学の問題ではあるけれども、電子資料作成の本質的な問題点ではなく、また分析・評価の手法として要求分析が当てはめられないことではない。例えば、ある資料を電子化する際に、その利用目的が分からない場合でも、もし資料そのもののに価値があると認められるのであれば、電子資料には二次資料としての価値がある。もし利用目的が分析できるのであれば、それらの要求に従い、後述するテキスト版における分析対象を規定すればよい。この事例一般にみられる注意点・課題としては、作られる資料の利用者と、製作の依頼主が異なる場合、例えば、図書館・文書館が資料製作を主導する場合に、要求分析が、手続き上は完遂されても、実質の利用者である研究者の要求が分析の対象とされないことがある。

4.2 モデル研究

科学の分野では、物理学の公式のような真理とはならないが、研究を実施するためには必要となる、真理に代わる仮想の実体 (最終的な発見の目標で、到達はできないが、それに近似できるもの) を、モデルと称している。物理学の公式のような真理は、理想的な実験や観察結果と常に一致するが、モデル研究では、モデルが実測値と一致することは叶わないので、そのモデルは実測値に近づけるため常に修正される対象となる。モデルは、実験や観察から想定され、これまでに得られていない結果をモデルから予測し、その予測度の良さが判定・検定され、それらの結果を踏まえてモデルは修正されてゆく。モデル研究における分析・評価の手法は、システム開発と同じく、現象 (要求) を分析してモデル (仕様・設計) を (仮に) 定義し、その当てはまりの良さ (評価) をくり返すことで、モデルの完成度を上げてゆく。

人文情報学では、(y) 非テキストを対象とした電子資料の作成において、この手法を採ることになる¹⁸。非テキスト媒体を電子化する AD 変換では、データ構造を決める際にモデルを規定する必要がある。エンタテインメントの分野では、このモデルは工夫点のひとつとして、使い勝手の良さで評価されることがある。人文情報学の場合、モデルの評価基準を、(1) モデルの対象となった現象との当てはまり度 (説明の良さ) とするのか、または、(2) 使い勝手の良さ、すなわち、現象を鑑賞・批評する際に便利であるかの、2つの基準が考えられる。(1) は、科学のモデル研究と似ているが、(2) は、人文学における人の批評活動そのものを比較の対象とすることから、人文学研究に直接還元できる成果を生む可能性はあるものの、分析・評価は難しくなる。従来の人文情報学では、モデル評価法を自らの評価法として導入することは多くなかったが、今後は増えることが予想される。例えば、(c) 電子資料/データから二次資料を作るとき、最終成果物の評価

¹⁵ この際、注意すべきは、モデル化の場合、単位化の基準は「工夫点」の一つとして、全体評価に紛れることになりがちであるが、モデル化の教科書にもよく示されるように、単位化の規定は、全体の結果に大きな影響を与えることから、本来は十分に配慮されるべきである。とりわけ、人文研究での利用を想定する場合には、この単位化そのものが人文学では大きな論点となることから、選択的に単位化ができる、または複数の単位化基準で全体の評価を比較するアプローチが、人文情報学では大切になってくる。

¹⁶ 得られたモデルやオブジェクトを、人文学研究でどう評価するのかは難しい。恐らく、この種のオブジェクトがまだ人文情報学の主たる成果物とはなっていない理由がここにあるのだろう。また、この評価法を採ることを止めて、従来の人文学における批評的な評価法、すなわちオブジェクトを「鑑賞」という傾向が、後述するように増えてきている現状も、これが背景としてあると考えられる。

¹⁷ 人、機械、ソフトを三体とする場合もある。

¹⁸ 本稿の区分では、非テキストを電子テキスト化するアプローチを人文情報学の活動対象から外している。

は、二次資料として存在そのものが評価されことに加えて、その基となるデータの検索・選択・抽出過程で想定した式・モデルを、得られた結果から評価することも可能である¹⁹。

4.3 テキスト版とファクシミリ版

資料を符号化する際に分析・評価対象や手法で問題とされたケースの多くは、テキスト資料を(電子)テキスト版として作成する場合である。また、(電子)ファクシミリ版を作成する際、内容検索の利便性を考え、同時にテキスト版を作成するときにも、同じ問題を抱えることになる。

ファクシミリ版とは、図書資料の一分類として、原本に代わり写真やコピー等の複写物としてある資料を指し、電子資料においては、当該資料の画像データが電子オブジェクトとなる。例えば、写本、古写真、古地図などはファクシミリ版オブジェクトとして電子化されることが多い。また何らかの理由でテキスト化が困難である活字印刷の書籍でも、ファクシミリ版は作られることがある。ファクシミリ版では、資料の意味内容を示す役割はメタデータが担うことになるため、メタデータは必須の付属データとなる。ファクシミリ版の作成とは、対象を静止画像として取り込む作業である。この作業そのものの分析・評価基準としては、既に撮影技術に関する多くの実験成果や標準の基準が存在していることから、電子化の現場で問題となるケースは少ない[11, 19]。ファクシミリ版の作成で問題となるのは、主にメタデータの項目で、これにはシステム開発の分析手法を採用ことになる。ファクシミリ版と同時にテキスト版が作成される場合、画像とテキストの連携が問題となるが、この分析にはテキスト版におけるそれが利用されることになる。

テキスト版とは、多くの場合、マークアップ言語によるメタ記述を含むテキストデータである²⁰。マークアップ言語そのものの選択は、重要ではあるものの、現時点ではXML以外の選択はHTML5しか考えられず、判断に悩むケースは少ないと思われる²¹。

テキスト版の作成では、何を符号化するのか、すなわち何をマークアップの単位(e.g. 要素, 属性)とし、それをどのように表現するのか、すなわちどのようなデータ構造(e.g. 文書構造)とするのが、現在まで続く人文情報学の主要な論点とされているが、この問題については、TEI[6]が示しているように、1) 何を単位化するか、すなわち何をマークアップするかは、テキスト種/ジャンルの研究コミュニティの要望で決まり、2) その表現方法、すなわち SGML/XML を使う場合には文書構造(syntax,

scheme) は、データの可換性を鑑みて規定されることが、分析の原則となっている。これらの詳細は、5章で扱う。ちなみに、近年、新しいスキーム研究の流れとして、従来の文書構造を模した木構造ではない、データ構造が、ISO[7, 8, 9]を中心に提案されている。現在進行している各プロジェクトがこれらの提案にどこまで追従するのかは分からないが、人文資料をグラフ構造で表現した電子資料の作成はこれまでにないことから、新たな実験・研究が必要となる。

4.4 Ontological Semantics

メタデータにアノテーションの記述が求められるようになると、どのような情報を第三の著者が記入すべきかは、新たな分析の対象となる。もし、要求分析が採用されるケース、例えば(i) テキスト版電子資料を作成する際には、アノテーションも同様に、利用者の要求に合わせて作成されることが考えられる。ところが、要求分析が採られないケース、例えば、(y) 非テキストを対象とした電子資料や、(c) 電子資料を元にする二次資料を作成する際には、アノテーションの内容を決めることが難しいこともある²²。従来は、(y) や (c) を作成する際には、アノテーションをメタデータに記録することは少なかったが、現在は、Semantic Web 研究などの様にアノテーションは積極的に記録されている。このケースで利用される、分析・評価の対象と手法検討の枠組みとして、情報理論の考え方と、人文学にある二次資料の価値の決め方を採ることができる。

情報理論は、記号の出現状況のみを分析するもので、記号の意味は処理の対象とはされない。情報理論のコンセプトに従えば、人の行動の痕跡として記号化されたリンクのみが観察対象の記号であり、リンクの意味を示すアノテーションは処理対象から外される。アノテーションをメタデータに記入するだけではなく、データ本体中の部分単位にリンクでもってアノテーションを付加したり、そのリンク自体に意味役割をアノテーションとして付加しても、これらは全て ontological semantics の処理対象とはならない。情報理論のコンセプトの延長にあるとされる ontological Semantics では、外形の記号のみを分析対象として、人の判断や解釈などの意味世界に関する記録は分析の対象とはしない²³。

このような ontological Semantics の考えに従うと、アノテーションそのものは、分析の対象からは外れることから、何をアノテーションとしても資料の評価には影響を与えないことにもなる²⁴。つまり、符号化の設計段階において、アノテーション項目の選別に悩む必要はなく、利用者が必要と判断した情報をそのまま入力することができる。このような場合、アノテーションは際限なく自由に記述される可能性もあるが、現実には、従来の人文学

¹⁹ すなわち、従来あった人文学と計算機科学の双方の評価方法を、同時に採ることになる。

²⁰ プレインテキストによる電子化で十分なケースでは、例えば、文字が問題とならない場合には、分析・評価基準で悩むことは少ないだろう。

²¹ いわゆるオーバーラップの問題を理由に、XML 以外のマークアップ言語を模索する研究もあるが、あまりよいアプローチではない。理由は、1) そもそも分析対象の整理が間違っている可能性がある、2) 線状性を利用しないデータ構造を利用することもできる、3) 抽象度または分類が異なるデータを共起記述させる理由を十分に検討していない可能性がある、4) 現在提案されている非 XML マークアップ言語では、記述の規則を元にした記述の階層についての論議が少なく、結果として、処理を難しくしている、などがある。

²² 例えば、発見型のアプローチでは、事前に要求を決めることができない。ちなみに、どの様に符号化されたのかという情報は、従来の電子資料でもメタデータとして記録されていることから、電子化の様子を振り返ることは従来のメタデータからもできる。

²³ 例えば、このようなスタンスを取る研究として、通称「Big Data」呼ばれる研究があるが、これは通称 Semantic Web と呼ばれる研究と異なり、意味を記号化する工程がないことから、情報理論の哲学に照らせば、この2つは本質的に異なる研究になっている。

²⁴ データ化されても、資料を汚すと評される情報にはならない。

で二次資料を決める際に採られてきた、特定研究コミュニティからの評価、すなわち人文研究者が有用と判断してきたことを基準とすることで、従来以上の評価基準は生まれないことになる。アノテーションは、書かれなくとも、書かれても良く、書かれる場合、その内容は人文学者の判断/求めて決められる。但し、メタデータやアノテーションそのものも電子テキスト/オブジェクトであるから、アノテーションの本文から何を単位化するのか、その構造をどうするのか、という、テキスト版の作成時に必要となる分析が必要となる(5章)。但し、先取りをすれば、アノテーションをアノテーションの対象とする記述を可能にするスキームは、残念ながら現状ではまだよいものが提案されていない²⁵。

まとめると、アノテーションは、利用者の要求に合わせて作られ、それが出来ない場合には、二次記録として(データを作成する)人文学者が必要と判断したものを全て記録することができ、それは他のアプローチを採る研究に悪影響を与えることはない。また、もし適したマークアップの判断がつかない場合には、符号化する必要もない²⁶。

5. テキスト版の分析・評価対象

テキスト資料から電子テキスト版(i)を作成する場合、マークアップ言語を使いマークアップ付きテキストを作成することが多い²⁷。この時、何をマークアップの単位とするのか、どのようなデータ構造で記述するのかを分析し、規定することは、人文情報学では長年課題のひとつとされてきた。この課題は、1) マークアップ言語の使い方についての論議と2) 符号化する資料の分析の2つに分けることができる。1) については、TEI[6]で示されているように、a) 単位は、研究コミュニティの要望で決まり、b) その単位の構造、すなわちスキームは、データの可換性で決めることが原則となっている。当該資料を利用する研究コミュニティが必要とするものが符号化される単位、すなわち要素と属性となり、その構造は、少なくとも当該コミュニティの中での可換性、または、要求として求められたコミュニティ内での可換性にあわせて規定されることになる²⁸。

SGML以降のマークアップ言語の応用では、資料の符号化をはじめにあたり、次の点が分析の対象とされてきた。1) 要素をつける単位、2) 属性として記される情報、3) 要素名または属性(属性名、属性値)のどちらで記するか、4) 要素の構造はどうするか、5) 処理をどのように想定するか(e.g. 実体、リンク、特定ソフトなど)。

5.1 メタ記述単位

マークアップ言語によるメタ記述は、要素名と属性(属性名、属性値)がある。これらメタ記述は、電子化されるテキストに対するアノテーション、すなわち、当該対象の著者とは異なる作者が記述したテキストである。要素名は、名前付与規則に抵触しない限り、アノテーションの作者(符号化する人)が自由に決めることができる。この特徴を利用して、アノテーションの作者は、要素名を意味記述として使うことにより、意味を規定する役割を期待することができる。但し、SGML/XMLの規格では、要素名に意味はない。これは、記述の規則²⁹でいうところの、表記と処理の関連性は定義されていないことを示すもので、人がそのように使うことができないということではない。但し、符号化する人以外とその意味が共有されないのであれば、SGML/XMLによる電子化の目的のひとつでもあるデータ共有を妨げることから、一般には、「意味を規定すると期待してはいけない」と警告されることになる³⁰。

属性(属性名、属性値)も、本来は要素名と同様に、符号化する人が自由に規定することができる。従って、意味表現については、要素名と同様の位置づけにある。また、属性は、ひとつの要素名に複数記述することができる。SGML/XMLの規格上、各属性の間には関連性はない。例えば、順序という関連性も存在しないことになる。また、属性は要素名に付加されるが、これが要素内容とどのように関連性を持つのか、例えば、属性が関連を持つ対象は、テキスト内容か、第一子要素か、全ての子要素か、要素名そのものかという関連性についての規定はない。一般には、要素名としてアノテートされた対象である要素内容全てと関連していると、人により解釈されることが期待される。但し、これも要素名と同様に、属性の表記と処理との関連性が定義されていないことから、関連性を自由に想定することができる。従って、データの可換性を重要視する場合には、特定の関連性を期待することは避けられる。特定コミュニティでの可換性を求めるのであれば、それをドキュメントとして記されることが期待されることになる³¹。

²⁵ 最新のISO24612[9]でも、この限界を克服できないでいる。この程度の高階の関係が表現できないと、例えば、2つの表現(リンクの属性を付加することなく)リンク(のみ)で関連性を意味として表現するスキームは規定できず、意味研究の障害となる。

²⁶ 要求が分からないのであるから、単位化する必要はないということ。つまり、アノテーション全体を示す要素の中に、テキスト内容として入れておくだけでよい。単位化の詳細は、次章にある。

²⁷ 日本では、そうとはいえないことは承知している。

²⁸ TEIは「特定研究コミュニティ(当該資料を使う人達)におけるデータの可換性を目的としている」もので、TEIへの誤解として良くある「TEIは人文資料一般のデータ処理を目的としている」ものではない。確かに、TEI P5には、そのような言及もある。また、TEIガイドラインP5以降、国内外でTEIを利用する研究者の発表では、スキーマが研究コミュニティからの要請で決められることなく、TEIを天下的に採用しているようなケースも報告されている。日本の研究者が、TEI P5からミスリードされないために、改めて歴史を確認すると、TEIガイドラインは、定義言語であるメタ言語/マークアップ言語をSGMLからXMLへと変更した過程で、処理を重視する編集方針が強まり、結果として現在のP5が作られているが、P5以降に加えられた各種処理が施されるのは、その処理対象となるデータのスキーマが確定した後の段階である。すなわち、データスキーマを策定する姿勢は、この20年何も変わるもの

はない(スキームに新たにクラスという概念が導入された変更点はある)。P3で推奨されている、スキームは研究コミュニティからの要請をもとに、データの可換性すなわち当該研究者間でデータ交換がしやすいデータ構造を決める、という方針は現在でも有効と考えるべきである(N.Ide and C.M.Sperberg-McQueen(1995)“The TEI”[6])。

²⁹ 紙面の都合でこれを解説した章を割愛してしまった。別稿に譲りますことをお許しください。

³⁰ まとめると、特定コミュニティ内で意味が共有できる(e.g. TEIの要素名)のであれば、意味を規定しても良い。この場合、特定コミュニティ外では、XMLが保障しない意味共有が、データ共有の障害となる。

³¹ TEIは、独自の属性を規定し、それをドキュメントにまとめている。ISO[7, 8]では、そのように独自に要素名や属性を規定する手段が規格化されている。

5.2 スキーム

SGML/XML では、要素名や属性 (属性名, 属性値) として単位化された情報を、まとめて要素という単位でパッケージ化し、これを構成要素としてデータ構造を作る。SGML/XML では、要素をダイク (Dyck) 言語 [5] としてテキストにデータ構造が付けられる。

SGML/XML では、要素における構造、すなわち要素名と属性 (属性名, 属性値) から成る構造と、要素が構成する構造の、2 種類が記述される。前者では、属性間には関連性が無いため、単純に、要素名, 属性名, 属性値のトリオが基本構造となる。後者では、木構造が構成されている。ちなみに、SGML/XML では、データファイルを一つの要素で単位化されることが求められ、その要素は木構造の根要素 (root element) とされる。

このような、要素における構造や、要素が構成する構造で表現されているデータ構造をまとめて、スキーム (scheme) という。スキームは、文書構造をマークアップ言語で表現するために作られたデータ構造である。SGML では、スキームを、SGML とは異なる記述規則を採用した DTD により定義する。XML では、スキームの定義は選択可能なものとなり、要求に合わせて構造のパターンを指定できるようになった。このような構造を指定する構造パターンは、データベースに習いスキーマ (schema) と呼ばれ、このパターンの定義方法はスキーマ言語と呼ばれる³²。

XML では、スキームが定義されているデータを valid XML data (妥当な XML データ)、スキームが定義されていないデータを well-formed XML data (整形形式の XML データ) として分けている。SGML データと、valid XML data は、定義されている (部分) データを、インスタンス (instance) と呼ぶ。ところが、well-formed XML data では、スキーム定義がないことから、その (部分) データは、厳密にはインスタンスとはいえない。残念ながらこれを示す用語が定義されていないことから、本稿ではこれを instantiations と表現する。

スキームは、実 XML データに記述されているデータ構造であるが、このスキームを抽象化する単位は、スキーマ言語が多様に存在するように、多様にある。例えば、要素のみ (すなわち、メタ記述のみ) で構成される記述データ³³は、マークアップデータのパーサを実装する際に、重要な役割を果たす³⁴。

5.3 モデル構造

SGML/XML におけるスキームは、実 SGML/XML データのデータ構造を指すだけでなく、何かしらの処理が施された後のデータにあるデータ構造を示すことができる。例えば、実体 (entity) や PI (Processing Instruction) は、何かしらの処理の結果としてあるデータを表現し、このデータ構造をスキームとすることがある。後者のデー

タ構造は、記述された記号の意味が解釈されて作られた結果であるから、厳密に言えば、モデル側のデータ構造である。例えば、このようなモデル側のデータ構造として、属性同士の参照関係 (IDREF-ID 関係) を解決し、いわゆるリンク構造をモデル上に想定したものがある [9]。XML の応用規格である RDF では、木構造によるモデルと、リンクを解釈することで示されるリンク構造のモデルが想定されている。

SGML/XML では、処理の局面を暗に含むモデル構造の論議は、多くが棚上げされたまま、今日に至っている³⁵。そのため、各 SGML/XML アプリケーション間では、このモデル構造の扱いが問題となり、複数 SGML/XML アプリケーションを同時に扱うことが困難である原因のひとつになっている³⁶。また、人文情報学でテキスト版電子資料を作成する場合、スキームとしてある木構造と、モデル上にあるリンク構造が、同時に扱われ、問題とされることがある。代表的な例としては、いわゆるオーバーラップと呼ばれる、1 つの XML データの中に文書構造と紙面を構成する印刷単位 (とその構造) とを同時に記述することが難しいという問題がある³⁷。

このような、記述の構造とモデルの構造との関連性が問題となるケースでは、問題の解決は、トレードオフを見つける作業に成らざるを得ず、より詳細な要求分析が必要となる。ちなみに、ISO から提案されている一連の規格群 [7, 8, 9] では、モデル上のデータ構造を積極的に活用するスキームを採用している³⁸。

5.4 まとめ

符号化される資料で、研究上の要請または資料の特性上必要と判断したものを、メタ記述単位、すなわち要素名と属性として単位化することになる。それらデータ単位の構造は、(1) 当該資料にある構造、例えば文書構造を反映したスキームとするか、(2) 符号化の要求に応じて、適時必要としたデータ構造をスキームとするか、(3) 場合によっては、むりに木構造を作る必要はなく、できるだけフラットなデータ構造のスキームにすることもある。また、スキームは、データの可換性を考慮して決められることから、(1)(2)(3) の全てで、少なくとも特定研究コミュニティでデータが可換であることを要求として検討する必要がある。従って、(4) TEI が推奨するスキームを全面的または部分的に採用することが考えられる³⁹。モデル上のデータ構造を必要とする場合は、研究上と資料の特性上からの要求に加えて、使用するソフトウェアま

³⁵ 例えば、SGML でいえば OMITTAG や PI の扱い、XML でいえば XML 宣言、名前空間宣言、実体処理など、きりが無い。

³⁶ この意味において、データ処理を規格化する HTML5 を、処理間の規定がない XML として XHTML 化することは、よい方向とは思えない。

³⁷ 例えば、新聞連載小説では 1 行の文字数が固定されることから、作者が特定文字数を文章のリズムを構成する要素として意識していた可能性があるとするれば、印刷単位である「行」を電子資料に記録しておくことは重要になる。このようなオーバーラップへの対応は、4.3 節にあるよう、人・資料・技術の総合的な分析を必要とするため、この解決が主たる研究テーマではない場合、検討はしない方がよい。

³⁸ 従って、各プロジェクトでトレードオフが検討された結果として複数のアーキテクチャが存在し、データ共有の障害となる可能性がある。現在、ISO ではテキストデータの可換性を高める為の規格群が精力的に検討されており、この動向は見守る価値がある。

³⁹ スキームの検討法については、拙著 [14] でも解説している。

³² 各スキーマ言語が示すスキームは同じではない。あるスキーム (scheme) は、あるスキーマ (schema) では表現できないこともある。従って、データ構造を表現するスキームと構造パターンを指定するスキーマは、厳密には分ける必要がある。

³³ 個人的には skeleton と呼んでいる [4]。

³⁴ この章では XML が作る情報単位を改めて整理した。これは [16] で一部を報告した科研費プロジェクトで使われたオリジナルのモデルを元にしている。このような試みとしては [17] がある。

でを想定して、システムの設計、いわばアーキテクチャを検討する必要がある。従って、電子二次資料の作成が目的であれば、この論議を避けた要求をまとめた方がよい。ちなみに、テキスト版を作成する際、表記法や利用方法 (e.g. 公開方法、保守方法) を、仕様として要求することがある。本稿で示す枠組みでは、これらは扱っていない⁴⁰。

6. 人文情報学の拡大に伴う補足

従来の born-digital オブジェクトと digitized オブジェクトの単純な区分では済まず、現在では電子資料を元に二次電子資料が作られてきている [2]。このような学術資料の製作では、特定コミュニティからの要求が分析・評価基準となる。従って、二次電子資料は研究活動の一環として作られることが可能となり、市場からの制作費回収を義務として負う必要が無くなることから、今後はより活発になることが予想される⁴¹。このような電子資料が、研究対象として一般化する中で、これらを鑑賞・批評の対象とする研究が人文情報学で盛んになっている [2, 10]。この活動では、計算機は電子資料を再生するためだけに使われ [12]、問題解決の道具としては使用されることなく、計算機科学や人間科学における問題解決法も採用されない [18]。このような研究活動の変化は人文学の出来事ではあるけれども、学際研究として生まれた人文情報学は、これに無関心でいるわけにはいかない。但し、本稿で提示した分析・評価基準の枠組みでは、このような人文情報学の活動は、検討の対象から外している。理由は、鑑賞・批評の目的を仕様としてまとめる手法が分からないこと、電子資料の部分へアノテーションを付ける技術が多様であることがある。

7. さいごに

本稿は、拙著 [15] に続いて人文情報学の応用知識を学ぶために用意したプロジェクトマネジメントの資料を再構成したものである。紙面の都合上、記述の理論、紙資料と電子資料の差異⁴²、メタデータ設計の分析事例などの項目を割愛しているが、分析・評価基準の枠組みの解説に必要な内容は出来るだけ収録した。新しく人文情報学に参加を希望する研究者の殆どが、アプローチの選択に悩むようであることから、その解消に必要な資料は今後も用意してゆきたい。情報処理学会でこのような資料が公開できたことに感謝を申し上げる。

参考文献

- [1] Association of Research Libraries, *SPEC Kit 326: Digital Humanities*, ARL, (2011)
- [2] D.M.Berry, *Understanding Digital Humanities*, Palgrave Macmillan, (2012)
- [3] D.J.Cohen and T.Scheinfeld eds., *Hacking the Academy*, The University of Michigan Press, (2013)
- [4] J.H.Coombs, A.H.Renear, and S.J.DeRose, “Markup Systems and The Future of Scholarly Text Processing”, *Communications of the ACM* Vol.30, No.30 Issue 11, ACM, (1987)
- [5] M.D.Davis, R.Sigal and E.J.Weyuker, *Computability, Complexity, and Languages*, Morgan Kaufmann, (1994)
- [6] N.Ide and J.Véronis eds, *Text Encoding Initiative*, Kluwer Academic Publishers, (1995)
- [7] ISO, *ISO/DIS 24610-1 Language Resource Management – Feature Structures – Part1: Feature Structure Representation*, ISO, (2006)
- [8] ISO, *Language resource management – Feature structures – Part 2: Feature system declaration*, ISO, (2011)
- [9] ISO 24612, *Language resource management – Linguistic annotation framework (LAF)*, ISO, (2012)
- [10] M.K.Gold ed., *Debates in the Digital Humanities*, University of Minnesota Press, (2012)
- [11] A.R.Kenney and O.Y.Rieger, *Moving Theory into Practice*, RLG, (2000)
- [12] L.Manovich, *Software Takes Command*, Bloomsbury Academic, (2013)
- [13] 大矢一志, 土屋俊, 「システムが決まらなければデータベースが出来ないというのは本当か」『第2回アートコミュニケーションフォーラム報告書』, 日本図書館協会, (2000)
- [14] 大矢一志, 「マークアップの課題を syntax から見た分類と解決のステップ」, 『TEI Day in Kyoto 2006 報告書』, 京都大学, (2006)
- [15] 大矢一志, 『人文情報学への招待』, 神奈川新聞社, (2011)
- [16] K.Ohya, “Corpus Sharing Strategy for Descriptive Linguistics”, JADH2012, (2012)
- [17] J. Vion-Dury, “Using RDFS/OWL to Ease Semantic Integration of Structured Documents”, *Proceedings of the 2013 ACM symposium on Document engineering*, ACM, (2013)
- [18] S.T.de Zepetnek ed., *Digital Humanities and the Study of Intermediality in Comparative Cultural Studies*, Purdue Scholarly Publishing Services, (2013)
- [19] C.Warwick, M.Terras and J.Nyhan ed., *Digital Humanities in Practice*, Facet Publishing, (2012)
- [20] A.W.Wymore, *Model-based Systems Engineering*, CRC Press, (1993)
- [21] D.M.Zorich, *A Survey of Digital Humanities Centers in the United States*, CLIR, (2008)

⁴⁰ 個人的には、これらは概論を超えた各論扱いのもので、研究対象そのものと考えられる。

⁴¹ 研究資金で電子資料が作られることは、反面、市場の資金から紙の学術資料が作られなく可能性を高める恐れがある。教育環境と研究成果の評価方法は再考されることになるのかもしれない [3]。

⁴² ハイパーリンクが批評的読解力に影響を与えるのか。