人文科学とコンピュータ 56－3
（２００２．１１．１５）

「文化財デジタルアーカイブにおけるシソーラス研究」
～インフラストラクチャーとしてのシソーラス～

早稲田大学大学院国際情報通信研究科（GITS）
早稲田大学文化遺産デジタルアーカイブ研究所
平尾大輔

シソーラスとは（thesaurus）とは語句を“意味”によって分類・配列した語彙集であり、分類語彙集という場合もある。文化公共施設デジタルアーカイブにおける言語情報検索シソーラスの考察を行う。文化財：文化対象に付されたキーワードの示す範囲、キーワードとの関連語の類似・対立・包括関係等などシステムを総括した研究。

「Research in the thesaurus in the field of digital archives for cultural properties」

Graduate School of Global Information and Telecommunication Studies
Waseda Institute of Museology ：Digital Archives and Conservation Science ／CEO
HIRAO Daisuke

Summary: “Thesaurus” is a collection of vocabulary, which classifies and arranges words according to ”meaning” and depending on the circumstances, it may mean a collection of synonym. We research in the thesaurus for searching for language information. So we consider the range of meaning attached to cultural properties : cultural objects and the relationship between keywords and related words, for example, synonym, opposition and inclusion.

１、はじめに

シソーラスとはインフラストラクチャーである。IT ファシリティーの社会整備進展は目覚ましく、高機能情報機器と超高速通信は急速な拡充を見せている。対する情報コンテンツも電子政府化で進められているが、もう一方の翼、市民生活や教育をはじめ各種産業基盤に必須の文化コンテンツデジタル化の遅れが著しい。その理由は多種考えられるが、システム構築の基礎となるシソーラスの設定の不在は最も大きな要因である。各種コンテンツが自動車や飛行機とすると、シソーラスは道路や飛行場に相当する。また、コンテンツ制作を相当する各地方自治体

や学術、教育機関にとって、インフラストラクチャー研究開発を個別に行うことは効率的とはいえず、高いクオリティーは望めない。研究開発体制の整備と社会的普及は急務と考える。

本稿ではデジタルアーカイブにおけるシソーラスに関し、以下の二点について考察を述べる。

A. 現行においてアーカイブがどの様に概念設定されるかの解析とそのシステムへの影響

B. 文化財：文化的対象に必要とされるシソーラスの特性

考察の根拠として、A は、近年内外を問わず文科公共施設や教育機関、企業における取り組みが増大しているデジタルアーカイブという領域に関し、基礎認識や表現概念の曖昧さが見受けられ、それがデジタルアーカイブのシステム構築のみならずシステムの社会運用にも不合理な状況をもたらしている点。そして、開発及び研究件数の増大に比例し、それらが振幅が大きくなってきていること。また、それらの状況が B に対し、主にシソーラス機能に必要とされる技術的要素や条件を不明瞭にする要因に繋がっている。

考察を通し、情報テクノロジーとしての具体的な研究開発要素とその領域をより明確化し効率化を計り、文化コンテンツを対象とした市民社会への多様な活用、産業分野への基礎資料応用及び学術領域の発展を促す事を主眼としている。本稿における文化財という用語は、日本における国や自治体指定文化財を指すものではなく、広く文化的対象（文中状況により併記）として使用している。国連の国際ミュージアム評議会（ICOM）の規約において、近年その範囲は文化的財（biens cultuels）から、人間とその環境との物的証拠（temoins mateeriels）として推移しており、価値の多様性や他分野との関連性を含めた概念として捉えられてきている。この傾向は文化を題材としたデジタルアーカイブにとっても、重要な意味を持つものと推察する。

尚、ミュージアム及びミュゼオロジーの語彙範囲には博物館、美術館、歴史博物館、古文書館の他、図書館等、広く文化公共施設、社会における文化機能も含まれる。

２、序論～初期設定～

デジタルアーカイブとはデジタルテクノロジーを用いた広義の公文書資料保存庫である。アーカイブ Archive の語源はラテン語ギリシャ語で public office、ギリシャ語の ahkhe は government “ 政府 ” の意。歴史的価値あるものを位置付け “ 守り conservation ”“ 公く（ひらく） pubulic managaement ” 事が核となる。（尚、WWW 制作のハイパーテキスト記述言語 HTML は米国公文書書式 SGML を簡略化したもの。HTML はアーカイブの申し子と言える）

シソーラスとは語句を “ 意味類似 ” によって分類、配列したもの。分類語彙集。１８５１年イギリスのロジェが刊行した辞典名に由来する。コンピューター以前の図書館においても、書籍検索のためのシステムは図書館学における分類学 clasifikeation 分野で研究開発は進められてきた。図書の検索をことばで行う場合、自由な自然語を利用すると、どうしても同一の意味を

持つが表記が異なるデータが漏れてしまう可能性がある。それを防ぐために考えられたのが、索引やカードシステムに代表される“統制語”という概念である。統制語を使用するということは、同義・類語の類いを一つだけそのカテゴリーにまとめるということであり、この考えをすすめるにあたってシソーラスは発達してきた。

現在コンピューターによる様々な検索システムが発達して来ているが、多くはこの図書館方式、シソーラスの概念を基としたものである。検索システムはシソーラスの申し子と言うことになる。

これに対して、ハードディスク等の記憶媒介が安くなり量的な制限がなくなって来たこと、また、CPU の高速化と安価安定普及により、全てのデータを取り込む検索を行う全文検索が発達してきた。

それぞれの特徴を示す

・（図書検索に見られる）統制語を使用しての検索～あるいは専門用語を使用しての検索～

長所：正しい言葉を使用すれば全てのデータにヒットする

短所：正しい言葉を知らないとヒットさせることが難しい

・（WEB などでよく見られる）自然語による全文検索～検索エンジン、サーチエンジン～

長所：統制された言葉、専門的な言葉を知らなくてもデータの中で検索語があればヒットする

短所：データの中に存在しない言葉ではヒットしない

特質としては、前者においては、自然語から統制語を導く事によって、後者は同義・類義語も検索対象とすることによって、検索率とそのクオリティはより高まる点があげられる。検索エンジンを始めとするシステムはインターネット出現により産業として大きく注目され開発が進められているが、機能の飛躍的向上の要として、シソーラスの重要度が認識されて来ている。

尚、デジタルの領域において使用されるシソーラスには、分類語彙辞典以外に検索機能（エンジン）を備えたものがあり、双方備わったものが、いずれかの呼び方で総称されるケースが見受けられる。

３、海外における取り組み

デジタルアーカイブにおけるシソーラス研究の代表的な取り組みとして、EU ヨーロッパ連合、イギリス、アメリカとカナダの例を挙げる。

フランス美術館修復研究センター（C2RMF）は、作品履歴、研究データ、作品状態調査書、修復前、途中、後詳細写真（直射光・射光線・紫外線・赤外線）、X 線等光学調査、物理科学分析調査、年代測定等のデータを対象とし、索引化の際あらゆる曖昧さを避けるために用語の標準化：シソーラス研究と構築が行われた。これは EU の NARCISSE ナルシス（Network of Art

Resarch Computer Image Systems in Europe）プロジェクトとして、八か国語に定義・翻訳された各分野専門辞典を含む総合シソーラスの新データベース管理システムによって進められた。用語とその定義を確定することにより、索引後の使用に当って“知識の共有化”が可能となり、美術の素材、製作技術、様式とその変化、保存修復等の特色を正確に示すことができるようになった。現在は CRISTAL プロジェクトと名称を変更して進行中であり、シソーラスを始め今後の研究を推進するための基礎フォーマットを形作るものとして期待されている。

これらの文化プロジェクトは、EU というまとまりで解るように、主に、対米勢力対策として欧州の地位向上を目し進められた。同時に、学術や教育のみならず、観光、窯業、繊維、ファッション、インテリア、建築、インダストリアルデザイン等のデータベースとして、欧州産業振興の意も十全に含んだものとなっている。

MDA(イギリス／ミュージアムドキュメント協会)がドキュメント全般の標準化を２０年間近く進めており、その中で、シソーラスやシソーラス構築のガイドラインを規定している。

ATT(The Art and Architectures Thesaurus)これはアメリカのゲッティ美術史情報プログラムと、カナダの CHIN カナダ文化財情報ネットワーク（Canadian hiiritage information network）と共同開発したもの。美術情報の書誌、ビジュアル・データベース、資料コレクションのドキュメンテーションで使われる美術・建築専門用語のシソーラスの基礎となっている。

以上の例で明らかなように、文化財：文化財対象のデータベース設計、デジタルアーカイブ構築において、先ずシソーラス研究開発がその要としてプロジェクトに組み入れられているのが状況である。

ただ、欧米文化圏はラテン語ギリシャ語から派生し、国、民族、地域、階層の言語が成立した経緯から、一般的知識人が国語辞書と同時に類語（類義語）辞書を併用して広く使用する土壌がある。このようなシソーラスそのものに対する認識の差も、日本において研究開発が遅れている要因として考えられる。

国内においては『デジタルアーカイブを設定している研究者は同じ分野の研究者を対処としており、他分野や一般は考えに無い』と云う言葉が聞かれるように、先のシソーラス認識をはじめ、専門家においても情報化への取り組みに大きな差がみられるのが現状である。

４、文化財：文化対象アーカイブにおける検索機能

文化財デジタルアーカイブ構築時の必須課題として、検索機能の向上があげられる。対象物に関する情報は文字や数字、音声、画像、映像等、多様な形態を持ち得るが、従来は情報機器、ハードディスクの容量や処理速度の条件から、それらを対象物の表題や任意登録番号のみで検索するデータベースが一般的だった。しかし文化財対象物につけられる表題や登録番号は、特

定分野の専門家や研究当事者のみが理解する専門用語であることが多く、それ以外の人が表題を正確に表記し検索する事は非常に難しいという問題点があった。対象物を任意カテゴリー別に整理してある事例や、複数の索引語を設定している事例に他分野あるいは一般の人が使用する類似表現や意味体系で探すことが大変困難であった。

視覚表現を主とする文化財など、画像を見ることによる検索が有用であるとして研究模索されているものも存在するが、そうしたデジタルアーカイブも含めて、言語（キーワード）による検索の自由度の利便性を追求した『シソーラス』の検討と導入が必要不可欠と推定される。データベースに全文検索機能を取り入れた場合は、キーワードに含まれる全ての情報を検索してくれるが、同義・類義の稀ワードが含まれていても、キーワードそのものが含まれていない情報は全て対象外なってしまう。又、一般的な類語辞書機能を導入し他場合は、辞書が文化財保存科学関連の専門用語を網羅していない為、日常的な言葉から専門用語によって記述された対象を探し出すことはできない。上記問題点を十全な要素項目としてデジタルアーカイブに導入するシソーラスシステムは総合的な考察が必要とされる。

５、シソーラス構築の設定～表現概念の影響～

シソーラス構築の概念としては、当該文化専門家及びその保存科学（保存修復）の各専門分野で使用されている用語の類型化のみならず、研究や保存修復作業に関わる他分野の専門家が使用する専門用語から、一般の人が興味を持ち検索に使用する語彙の対応まで、幅広い領域を網羅するプロジェクト体制が必要とされる。これは各専門領域の総合用語集を作る事とは本質的に異なり、単なる用語の羅列や数値的な関係性を明らかにしただけでは機能しない。上記環境を包括した“立体的な意味構造”を研究～整理統合と構築～することによって初めて、それらと類似・対立・包括関係にある語句によって検索をかけること、又、利便性の高い索引機能を盛り込むことが可能となる。

・ 構成概略～関係各領域の用語を調査し、構造化、関係性を付加し発展的に統合～

A. 同義語（各領域のカタカナ語、外国語、古語、同音異義語、略語、通称、習慣的造語、職人語、地域語、慣用句関連、それぞれの表記の揺れ））狭義語、広義語、反義語関連、関連語、バックボーンとなる専門領域語

B. 言語のレイヤリング（階層関係）レイヤーリンク（関連関係）言語の持つ別種の意味概念、オントロジー～存在と本体、知識・語彙・概念とそれらの関係を明確にしたもの～

C. 付加機能考察（語末尾一致、各種学習機能、差別語や科学倫理名土曜後セキュリティ機能）・構成進行

1、文化対象当該専門分野と保存科学専門分野、また、置かれている・保存されている環境、付属するドキュメンテーション、二次資料分野に対する用語研究

2、１の領域に関連専門分野に対する用語研究～平面的、重・複層的、立体的、横断的階層構造～

3、１、２の領域と一般用語との関連に必要とされる用語研究

＊事前に関連研究の調査整備を行い、学術や職能・方法など専門とされる用語の体系様式を一定のテンプレート化する。これを各種領域に当てはめて開発の推進を計る。

＊図書と文化財～一般的に考えられる美術品等～の差は、記号情報と物理的存在、複数と唯一・単数、（市場）価値にある。後者は唯一の物理的対象を“守る”保存科学の領域が、その物の文化領域と共に存在し必須の要素とされている。これらの状況を充分考察したストラクチャーでなくてはならない。

・応用カテゴリーのパターン

ケース１、デジタルアーカイブ実装～アーカイブ機構自体にシソーラスを組み合わせる。現在インターネットデータ検索ヤフーや Google がこれに当り、公立図書館システムも該当。

ケース２、ネットワーク～ユーザーが検索を行う際、主体的にネットワークシステムにてシソーラスシステムを利用する。一部のネット辞書が類似。欧米においてはこのシステムが存在する。

ケース３、ユーザー実装～ユーザー自体がシソーラスソフトをパッケージとして用意。一般、医療、工業規格シソーラス等は在り。又は標準 OS やワープロソフトの様に PC に組み込まれる。

＊ケース１、２は専門用語の付加更新、修正等が容易な利点がある。言語は生き物として学問の進展や社会変化により変化する。社会への普及（各自治体、学術機関、教育機関、企業暖帯、一般市民）状況を充分考察し、システムを構築する必要が在る。

コンピュータ用語におけるアーカイブは、複数のファイルを一つにまとめることとして簡便に使用されている。時間の属性により、日常の乗用ファイルと長期保存用ないしバックアップ用とする考えがあり、主にハードディスクから別のメディアに保存されたものを指す。通常その過程により一定のフォーマット化、圧縮と展開（解凍）が行われ、これらを行うソフトウェアをアーカイバーという。

コンピューター用語と一般社会用語が違う例は数多く見受けられるが、デジタルアーカイブはそれ自体コンピューターを媒介としたものである為、交錯した表現が多く見受けられる。上

記概念の延長に在るものと推定される内部コンテンツの一部、文化を対象とした CG や VR の実験・表現作品データ単体に用語が付されるケースが多い。（文化対象でない場合は使われることが無いようである）その多くが一般に開かれたものとは言えず、それを目したものでもない。コンピューター、デジタルの特質であるインタラクティビティ性のない一時系列の番組構成となっている。

パブリックな情報としての文化コンテンツ製作に対し、このような基礎認識の差や表現概念の曖昧さは、アーカイブのシステム構築のみならずシステムの社会運用にも不合理であり、概念の啓蒙普及を含めた対策が必要であるものと考える。

６、おわりに

文化を対象としたデジタルアーカイブにおけるシソーラス構築は、その社会的に与える影響の高さ、必然性は従来の一国辞書に匹敵するものと推察する。具体的な科学技術、経済産業の進行にもダイレクトに連動し、地域社会、市民の家庭生活の下支えとしての役を担うものである。提示された諸問題に関しても、文化のロジックのみで語る事なく、社会機能や経済産業の実効性を配し総合的な視点・プロジェクトで解を導いていきたい。

本論の考察は文化財シソーラスについて多角的な考察を行った概説である。各論の進展は研究所で継続し、随時報告を行う。御意見、関連研究の連絡等を願う。

（dhirao@mn.waseda.ac.jp）