用語集

本文中に出てくる専門用語や、補足が必要な用語を解説します。

あ

青空文庫　著作権が消滅した作品や著者が許諾した作品のテキストを公開しているインターネット上の電子図書館。編集者の富田倫生（とみたみちお）らが発起人となって 1998 年に活動を開始し、2019 年時点では 14,000 点以上の作品が青空文庫の Web サイト（https://www.aozora.gr.jp/）上で公開されている。

か

コンピュータビジョン　コンピューターを利用した画像認識技術およびそれを研究する分野の総称。

さ

時間解像度　連続した存在である時間を、コンピューター上の離散的なデータとしてどれだけ細密に表現できるかを示す精度のこと。

時間情報システムHuTime　HuTime プロジェクトが開発する時間情報解析ソフトウエア。年表や時系列グラフの表示、時間に基づくデータの抽出など、時間情報の処理や分析に関わる多彩な機能を提供する（http://www.hutime.jp/）。

射影変換行列　画像に対する幾何学的変換のひとつ。射影変換を適用することで、例えば横方向から撮影した紙面の画像を、真正面から撮影したように補正することができる。射影変換行列とは、画像データを行列データとみなした際に、射影変換の操作を表現する行列のこと。

セマンティックギャップ　現実世界において人間が理解する意味内容（セマンティクス）と、コンピューター上で世界をモデル化したデータとの間に存在する大きな差異（ギャップ）のこと。例えばデジタルカメラで撮影した写真に対して、人間は容易にその内容を理解し言葉（意味）で表現できるが、コンピューターはその内容をピクセル列のデータとして分析するため、人間

と同様の意味的記述を与えることにはさまざまな技術的困難がともなう。

セマンティックWeb技術　Webページの意味（セマンティクス）をコンピューターに理解可能なかたちで記述することを目的にした技術の総称。World Wide Webの発明者であるティム・バーナーズ＝リーによって提唱された。

た

ダブリンコア（Dublin Core）　デジタルデータのメタデータを記述するための語彙のセット。1995年に米国オハイオ州のダブリンで開催されたワークショップで提案された。"Title" や "Creator" など、メタデータを記述する15種類の基本語彙によって構成されている（http://dublincore.org/）。

特徴量ベクトル　画像や映像、テキストなどのデータを要約する特徴量（feature value）をベクトル表現したもの。画像認識でよく用いられる特徴量にはSIFT、SURFなどがある。特徴量をベクトル化することで、データ間の類似度を計算したり、似た特徴を持つデータをクラスタリングしたりするなどの操作が可能になる。

は

ハミング距離　情報理論の用語で、同じ文字数からなるふたつの文字列の中で、対応する位置にある異なる文字の個数のこと。例えば「11011」と「10001」という文字列のハミング距離は2である（2番目と4番目の文字が異なる）。

パラレルコーパス　ひとつの事柄について記述された言語や表記などが異なるふたつ以上のテクストを、文単位もしくは段落単位で対応させて構築したコーパス。

ピア・プロダクション（peer production）　互いに対等な個人が構成するコミュニティの活動を通じて、何らかの製品やサービスを開発すること。ハーバード・ロースクールの法学者ヨハイ・ベンクラーがインターネット時代の新しい知的生産の形態として提唱した。LinuxやWikipediaはピア・プロダクションの代表としてあげられる。マスコラボレーション（mass collaboration）とも。

分散型コラボレーション（distributed collaboration）
インターネットを駆使することで、多数の人びとが場所的制約にとらわれず特定の目的の達成のために協力すること。

A

API Application Protocol Interface. あるプログラムの機能を外部のプログラムから利用するために用意されたインターフェイス（窓口）のこと。

ArcGIS ESRI 社が開発・販売している地理情報システム（GIS）ソフトウエア。地理情報を収集、整理、管理、解析、伝達、配布するためのさまざまな機能を実装している（https://www.esrij.com/products/arcgis/）。

Awesome-IIIF IIIF に関連するツールやソフトウエアなどの情報をまとめたレポジトリ（情報の集積所）。ソースコード共有サービスの GitHub 上で公開されている（https://github.com/IIIF/awesome-iiif）。

B

BRIEF Binary Robust Independent Elementary Features. 画像データの特徴を記述する手法のひとつ。浮動小数点数ではなく二値コード列を使って特徴を記述する。二値コード列のハミング距離は高速に計算可能であるため、画像データ間のマッチングを高速に実行できる。

byobu.exe 2001 年に歴博が開発した超大画像自在閲覧システム。Windows 上で動くアプリケーション。2000 年に開発した「超拡大！江戸図屏風」を汎用化したもの。DeepZoom と同様に、倍率ごとに用意されたタイル画像を用いて、高精細画像を表示する。

byobu32x.ocx 2007 年に歴博が開発した超大画像自在閲覧システム。Internet Explorer 上の Active X プラグインとして動作する。歴博総合展示第 3 展示室において高精細画像を含むデジタルコンテンツを提供するために開発した。画面レイアウトをかなり自由にカスタマイズすることができる。高精細画像のデータ形式は byobu.exe と同一。

C

core Builder　メリーランド大学において開発されている XML の編集ツール。特に TEI 準拠のマークアップを Web ブラウザ上でメニュー選択操作で行うことができる（https://github.com/raffazizzi/coreBuilder）。

Creative Commons(CC)　クリエイティブ・コモンズは、クリエイティブ・コモンズ・ライセンス（CC ライセンス）を提供している国際的非営利組織とそのプロジェクトの総称。CC ライセンスは、著作権のあるコンテンツを新たなかたちで利用させるためのツール。CC0 以下合計 7 種類のライセンスが提供されている（https://creativecommons.org/）。

CWRC writer　Canadian Writing Research Collaboratory（CWRC）が開発している WYSIWYG（What You See Is What You Get）の XML エディター。Web ブラウザ上で XML の文書を編集することが可能（https://github.com/cwrc/CWRC-WriterBase）。

D

DCMI Metadata Terms　ごく基本的なメタデータしか表現できないダブリンコアを拡張するために提案されたメタデータ記述のための語彙セット。ダブリンコアの語彙も含めた 55 種類の語彙によって構成される。

DeepZoom　Microsoft によって開発された、Web ブラウザ上で高精細画像をスムーズに表示するための技術。画像を倍率ごとにタイル状に分割することで、ネットワーク通信量を抑えながら高精細画像の表示を可能にしている（https://www.microsoft.com/silverlight/deep-zoom/）。

DOI　デジタルオブジェクト識別子（Digital Object Identifier）。Web 上のデジタルデータに与えられる識別子。この識別子を付与することで、ユーザーと提供者の間に DOI ディレクトリを経由させることができ、それによりデータの URL が変更されてリンク切れになるなどの事態を防ぐことができる。

E

EAD　符号化記録史料記述（Encoded Archival Description）。アーカイブズ資料の目録記述を電子符号化する方法の国際標準。XML・SGML を使用して資

料のメタデータを記述する（https://www.loc.gov/ead/）。

F

FAST　Features from Accelerated Segment Test. 画像データ中の特徴的な箇所を検出するアルゴリズムのひとつ。与えられた画像に含まれるコーナーを高速に検出することができる。

FOAF　Friend of a Friend. 人とその交友関係についての諸情報を機械可読形式で記述することを目指す実験的なプロジェクト。http://www.foaf-project.org で運営されており、RDF を含むセマンティック Web の基礎技術が用いられている。

G

GitHub　分散ソースコード管理システムである git をベースにしたソースコード共有サービス。Linux を含むさまざまなオープンソース・プロジェクトが GitHub を利用している。2018 年に Microsoft 社によって買収された（https://github.co.jp/）。

H

HuTime暦変換サービス　HuTime プロジェクトが提供する機能のひとつ。和暦や西暦など、異なる暦で示された時間情報を相互に変換することができる（http://www.hutime.jp/basicdata/calendar/form.html）。

I

IIIF　International Image Interoperability Framework. デジタルアーカイブにおいて公開される画像にアクセスするための標準的 API を定める国際標準。デジタルアーカイブの画像資料は、公開機関によってバラバラな形式で提供されていたが、2011 年に大英図書館やスタンフォード大学などの共同作業を通じて、画像データに対する標準的なアクセスを定める IIIF のバージョン 1.0 が公開された。画像データへのアクセス手段を与える Image API、書誌データやアノテーションも含めたデータ公開形式を定める Presentation API など、2019 年時点で 4 つの API によって構成されている（https://iiif.io/）。

IRG Ideographic Rapporteur Group. ISO/IEC 10646 および Unicode への漢字の追加に対して検討を行う専門家のグループ。中国、日本、韓国を含む漢字使用国から招待された専門家により構成される。

ISAD(G) 国際標準記録史料記述（General International Standard Archival Description）。アーカイブズ資料目録記述の国際標準であり、第 1 版は 1994 年に国際文書館評議会（International Council on Archives; ICA）の記述標準特別委員会にて採択された。現行の第 2 版（ISAD(G)2nd）は 2000 年 9 月に行われた ICA 国際会議にて採択された（https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition）。

ISO/IEC 10646 電子符号化方式や符号化文字集合などを定める文字コードの国際標準規格。業界規格である Unicode と互換性を持つ。

J

JSONフォーマット JavaScript Object Notation. プログラミング言語の JavaScript でデータをテキスト表現する際に利用されるフォーマット。Web 上のデータ交換フォーマットとして XML と並んで広く利用される。

K

KMNIST 2018 年に人文学オープンデータ共同利用センター（CODH）が公開した、「くずし字」の文字画像データセット（Kuzushiji-MNIST）。MNIST と互換性のある形式で公開されており、機械学習研究に容易に利用することができる（http://codh.rois.ac.jp/kmnist/）。

L

Linked Open Data 構造化されたデータ同士を URI を介して「リンク」させたデータを Linked Data（LD）と呼ぶ。Linked Open Data（LOD）とは、Linked Data のうち、Creative Commons などのオープンライセンスで提供されるデータを指す（http://linkeddata.org/）。

M

MARC 機械可読目録（MAchine-Readable Cataloging）。図書館資料の目録を電

子化するための国際的標準フォーマット。

Meidawiki　ユーザーによるWebページの内容編集を可能にするウィキシステムのひとつ。Wikipediaおよびその姉妹プロジェクトで利用されている（https://www.mediawiki.org）。

Mirador　スタンフォード大学を中心に開発されているオープンソースのIIIF画像ビューワー。https://projectmirador.orgで公開されている。

MNISTデータセット　機械学習の分野で画像認識のタスクに用いられるデータセットのひとつ。7万枚の手書き文字の数字画像から構成される（http://codh.rois.ac.jp/kmnist/）。

O

OAISモデル　Reference Model for an Open Archival Information System. コンテンツに合わせて表現情報・コンテクスト情報・来歴情報・不変性情報などを同時に保存してひとつの情報パッケージとして理解する。そしてこの情報パッケージを流通の段階ごとに制御し、変更の情報やコンテンツの関連情報を同時に記録することを通じて、データを長く保存することを目指したモデル。

OCR　光学文字認識（Optical Character Recognition）。印刷された文書をスキャンし、そこに書かれている文字を電子テキストに変換する技術およびソフトウエアのこと。

Omeka　図書館、博物館、美術館などのWeb展示を作成するためのコンテンツマネジメントシステム（CMS）。https://omeka.orgで公開されている。

Open Annotation　Web上のリソースにアノテーションを付与する方式やそのデータモデルについて検討を行っていた、標準化団体W3C下のコミュニティ・グループ。2013年に活動を停止し、その活動は同じくW3C下のWeb Annotationワーキンググループに引き継がれた。

OpenCV　コンピュータビジョンの分野で広く使用されているオープンソースのライブラリ（ライブラリとは、頻繁に利用されるプログラム中の処理をパッ

ケージし、共有可能にしたもの）。画像認識分野で使用されるさまざまなアルゴリズムが標準機能として実装されている。

OpenSeadragon DeepZoom 形式の画像ファイルの表示に対応した JavaScript ライブラリ。Microsoft 社によって開発されていたがオープンソース化された（https://openseadragon.github.io/）。

Oxygen XML Editor Syncro Soft 社によって開発されている高機能な XML 編集用ソフトウエア。TEI テクストの編集を支援するさまざまな機能を提供しており、TEI コミュニティでは広く使用されている（https://www.oxygenxml.com/）。

P

Presentation API IIIIF を構成する API のひとつ。画像資料の公開形式を「マニフェスト」と呼ばれる JSON-LD 形式のファイルによって指定する（https://iiif.io/api/presentation/2.1/）。

Q

QGIS 地理情報システム（GIS）ソフトウエアのひとつで、地理情報の閲覧、編集、分析が可能。オープンソースのソフトウエアであり、無償で利用することができる（https://www.qgis.org/）。

R

RANSAC RANdom SAmple Consensus. 与えられた観測値に外れ値が含まれる可能性を考慮し、その影響を最小限に抑えるための「ロバスト推定」を実現するアルゴリズムのひとつ。

RDF Resource Description Framework. Web 上のメタデータを記述するために用いられる汎用データモデル。主語－述語－目的語の組である「トリプル」によってメタデータを表現する。セマンティック Web の基礎技術のひとつ。RDF/XML、TTL、JSON-LD、Linked.art JSON-LD、KML、GeoJSON、IIIF Manifest という計 8 種類のデータ形式

Rights Statement DPLA と Europeana によって作成された権利のあ

り方を示す表記。CC が著作権のライセンスを示すのに対し、Rightsstatements は権利がどのような状態になっているかを端的に示すものとして用いられる。著作権の有無（不明を含む）、それ以外の権利制限の有無などを示すことができる（Rights Statement.org）。

S

Script Encoding Initiative　少数民族が使用するマイナーな文字や、歴史的な書記体系で使用されていた文字の Unicode への登録をサポートする団体。カリフォルニア大学バークレー校の言語学部を拠点として活動している（http://www.linguistics.berkeley.edu/sei/index.html）。

T

TAPAS　TEI テクストを保存・公開するための共用レポジトリ。TEI コンソーシアムや米国博物館･図書館サービス機構（IMLS）、全米人文学基金（NEH）などの助成のもと運営されている（http://www.tapasproject.org）。

TEI　Text Encoding Initiative. 文学作品や歴史資料などの人文学資料を XML でエンコード（符号化）する際のガイドラインを策定する団体およびこの団体によるガイドライン（https://tei-c.org/）。

TEILib　Best Practices for TEI in Libraries. 図書館において TEI 準拠のテキストデータを作成するためのガイドライン。

Transcribe Bentham　功利主義の提唱者として知られる哲学者 J. ベンサム（1748-1832）による 60,000 ページに及ぶ全集未収録の遺稿を、ボランティアの手によりオンラインで文字起こしするプロジェクト。ユニバーシティー・カレッジ・ロンドン（UCL）によって運営されている。

U

Unicode　文字の電子符号化方式や符号化文字集合を定めた文字コードの業界規格。日本語を含む世界中の文字体系に対応しており、2019 年時点の最新バージョンである Unicdoe 12.0 は、150 言語にわたる 137,993 文字をカバーしている。Unicode の管理運営はユニコード・コンソーシアムによっ

て担われている。

URI　Universal Resource Identifier. インターネット上に存在するリソース（資源）を指し示すための識別子。リソースの「場所」を指し示す URL（Universal Resource Locator）の概念を拡張したもの。

X

XML　eXtendsiblee Markup Language. 任意の用途について拡張可能なマークアップ言語。Web 上のデータ交換の標準フォーマットとして広く利用されている。その前身である SGML からの移行を目的として開発され、1998 年に最初の仕様が策定された。

Z

Zoomify　Zoomify 社が提供する高精細画像を Web ブラウザ上でスムーズに表示するための JavaScript ライブラリ。簡便なビューワーは無料で使える。画像の作成環境やカスタマイズ可能なソースコード、技術サポートなどの提供は有料。PhotoShop は Zoomfy 形式画像の作成機能を持っており、高精細画像を手軽に作成することができる（http://www.zoomify.com/）。