データエンジニアリング領域テキスト

（プロトタイプ版）

本テキストは、文部科学省の生涯学習振興事業委託費による委託事業として、学校法人河原学園 河原電子ビジネス専門学校が実施した平成30年度「専修学校による地域産業中核的人材養成事業」の成果物です。

1. コンピュータの基礎

1.1. データとコンピュータ

現在、私たちは、膨大な量のデータに囲まれて生活しています。そして、日常生活や業務において、これらのデータを適切に利用して役立つ情報を得るためには、コンピュータが不可欠です。

近年では、スマートフォンの普及や計測技術、通信技術の発達に伴い、生み出されるデータの量は飛躍的に増えています。ビッグデータと呼ばれるこれらの大規模なデータを、コンピュータを利用して処理し、価値を引き出すための手法は「データサイエンス」と呼ばれます。

ここでは、データサイエンスの観点から、社会や企業活動におけるコンピュータの利用がどのような考えや概念にもとづくものかを理解しましょう。

1.1.1. データ、情報、ファイルとは何か

「データ」とは、さまざまな計測や観察によって、主に数値の形で得られる内容のことをいいます。たとえば、買い物をした場合にレジで受け取るレシートを例に挙げましょう。レシートに印刷されている個々の商品の値段や数量はデータです。現在、多くの販売店では、商品のバーコードを読んで、「いつ」、「どの商品が」、「いくつ」売れたのかをリアルタイムで記録する POS（Point Of Sales）システムを導入しています。この POS システムで蓄積された記録もデータに当たります。

一方、データからその意味を読み取り、目的を考慮してそこから引き出された内容を「情報」といいます。毎日の買い物で受け取ったレシートで卵の値段を比較すれば、「卵が値上がりしている」といった情報を得ることができます。また、販売店で POS システムの記録を集計すれば、「時間帯や曜日によってどの商品がどれくらい売れているか」がわかります。この情報を元にして、店の棚に並ぶ商品に過不足がないよう、在庫を適切に管理したり、入荷の計画をいち早く立てることができるようになります。

このように、データ自体は、客観的な事実を表したもので、さまざまな目的に用いられる原材料に過ぎません。データサイエンスの観点では、原材料であるデータを適切に分析して、そこから有用な情報を引き出すことが重要になります。

現在では、収集したデータの内容を処理、分析する作業はコンピュータで行うのが主流です。コンピュータ上では、取り込まれたあらゆるデータは「ファイル」という最小単位で保存されます（「1.3.1 ファイルとは」参照）。コンピュータを使って、ファイルを正しく扱うことは、データ分析の観点からも不可欠です。

1.1.2. データ処理、データ分析のプロセス

(1)　Excel で表形式にデータをまとめる

データには、学校における学力テストの成績、商品やサービスについてのアンケート調査、世帯の収入や人数等の国が作成する統計データなど、多種多様な内容があります。また、コンピュータに入力する、紙媒体に記入したものを回収する、スマートフォンや専用の読み取り端末などを使うというように、その収集手段もさまざまです。

収集したデータには、点数、人数、数量、金額、測定値といった数値を含む場合がほとんどです。数値が入ったデータは、表計算ソフトの Excel を使えば、基本的な集計を行ったり、グラフを表示して可視化することにより、傾向を把握できるようになります。

そこで、ここでは Excel を使ったデータ処理の基本的な流れを紹介します。ここでは、とある飲料メーカーの売上データを例に挙げて説明します。

企業や店舗の売上データは、「日付」、「顧客名」、「商品名」、「数量」といった項目から構成されます。

エクセルで新規ファイルを作成したら、まずシートの先頭行に必要な項目見出しを並べ、それぞれの列に該当する内容を入力します。たとえば、「日付」の列には販売した日付を、「顧客名」の列には販売先の会社名を入力します。さらに、1 件の売上データは 1 行にまとめて入力します。

こうして出来上がったのが図 1 の表です。データをこのような表にまとめ、Excel ファイルとして保存しておけば、第 3 章で紹介する Excel の機能を使って集計や分析、グラフの作成などができるようになります。

	A	B	C	D	E	F	G	H	I	J	K
1	NO	日付	顧客名	支社名	商品コード	商品名	分類	単価	数量	金額	
2	1	2018/1/7	ミムラ出版	本社	E1001	おいしい水α	その他	820	120	98,400	
3	2	2018/1/7	ミムラ出版	本社	E1002	熱々ポタージュ	その他	1,500	75	112,500	
4	3	2018/1/7	加藤システム	本社	E1001	おいしい水α	その他	820	150	123,000	
5	4	2018/1/7	加藤システム	本社	E1003	カップ麺セット	その他	1,800	150	270,000	
6	5	2018/1/8	加藤システム	本社	C1003	コーヒー無糖	コーヒー	2,000	450	900,000	
7	6	2018/1/8	東田フーズ	新宿支社	T1001	煎茶	お茶	1,170	60	70,200	
8	7	2018/1/8	東田フーズ	新宿支社	T1003	朝の紅茶	お茶	1,800	150	270,000	
9	8	2018/1/8	安藤不動産	新宿支社	E1001	おいしい水α	その他	820	90	73,800	
10	9	2018/1/8	安藤不動産	新宿支社	E1003	カップ麺セット	その他	1,800	75	135,000	
11	10	2018/1/9	安藤不動産	新宿支社	E1004	ミルクココア	その他	1,300	15	19,500	
12	11	2018/1/9	春日部トラベル	さいたま支社	E1001	おいしい水α	その他	820	90	73,800	
13	12	2018/1/9	春日部トラベル	さいたま支社	E1004	ミルクココア	その他	1,300	60	78,000	
14	13	2018/1/9	柿本食品	さいたま支社	C1001	ドリップコーヒー	コーヒー	2,150	300	645,000	
15	14	2018/1/10	柿本食品	さいたま支社	C1002	カフェオーレ	コーヒー	1,700	300	510,000	
16	15	2018/1/10	ハウジング田中	前橋支社	T1001	煎茶	お茶	1,170	105	122,850	
17	16	2018/1/10	ハウジング田中	前橋支社	C1001	ドリップコーヒー	コーヒー	2,150	75	161,250	
18	17	2018/1/10	ハウジング田中	前橋支社	C1003	コーヒー無糖	コーヒー	2,000	75	150,000	
19	18	2018/1/10	デザインハルタ	浦安支社	E1001	おいしい水α	その他	820	150	123,000	

＞＞画面 1-1-1

図 1　Excel に入力された売上データ

(2)　データクリーニングで事前にデータを整える

実際のデータ処理では、収集したデータはそのままでは分析には使えません。図１のような表の形に整理するまでに、多くの手間と時間をかけて修正作業を行うのが一般的です。

このデータの前処理作業は「データクレンジング」あるいは「データクリーニング」と呼ばれます。データクリーニングとは、その名前の通り、集計や分析ができる状態になるまでデータの不備をなくして整備する作業のことです。

下の図 2 の表には、①～③のような不備があるため、そのままでは Excel の機能を使って分析することができません。それぞれ、どのようにデータを修正すればいいのかを説明します。

①表現にばらつきがある

F 列の商品名には、同じ商品なのに「カップめんセット（F5 セル)」と「カップ麺セット（F10 セル)」という 2 通りの表現が見られます。ひらがなと漢字の表現が混在しているためです。このまま Excel で集計すると、この２つは別々の商品として集計されてしまいます。

同じ内容のデータは表現も完全に一致していなければ、正しい分析をすることはできません。漢字とひらがな、カタカナの違いや、記号の全角文字と半角文字など（これを「表記の揺れ」と言います）は統一しておきましょう。

②異常値、外れ値がある

I10 セルには販売数が「6518」と入力されていますが、これは正しい値でしょうか。同一の商品が一度にこれほど大量に販売されることは現実にはあり得ないと思われます。

このように、明らかに異常と思われる値や、極端に大きすぎたり小さすぎたりする「外れ値」がある場合は、それが入力ミスなのか実際の値であるのかを確認する必要があります。

③入力されていない項目がある

C4 セルの顧客名が空欄になっています。空欄のセルには、該当する内容がないため空欄にしているのか、これから何かを入力する未入力の状態なのか、あるいは操作を誤ってデータを削除してしまったのか、さまざまな状態が考えられます。こちらも詳細を確認し、必要に応じて値を補って入力します。

	A	B	C	D	E	F	G	H	I	J	K
1	NO	日付	顧客名	支社名	商品コード	商品名	分類	単価	数量	金額	
2	1	2018/1/7	ミムラ出版	本社	E1001	おいしい水α	その他	820	120	98,400	
3	2	2018/1/7	ミムラ出版	本社	E1002	熱々ポタージュ	その他	1,500	75	112,500	
4	3	2018/1/7		本社	E1001	おいしい水α	その他	820	150	123,000	
5	4	2018/1/7	加藤システム	本社	E1003	カップめんセット	その他	1,800	150	270,000	
6	5	2018/1/8	加藤システム	本社	C1003	コーヒー無糖	コーヒー	2,000	450	900,000	
7	6	2018/1/8	東田フーズ	新宿支社	T1001	煎茶	お茶	1,170	60	70,200	
8	7	2018/1/8	東田フーズ	新宿支社	T1003	朝の紅茶	お茶	1,800	150	270,000	
9	8	2018/1/8	安藤不動産	新宿支社	E1001	おいしい水α	その他	820	90	73,800	
10	9	2018/1/8	安藤不動産	新宿支社	E1003	カップ麺セット	その他	1,800	6518	11,732,400	
11	10	2018/1/9	安藤不動産	新宿支社	E1004	ミルクココア	その他	1,300	15	19,500	
12	11	2018/1/9	春日部トラベル	さいたま支社	E1001	おいしい水α	その他	820	90	73,800	
13	12	2018/1/9	春日部トラベル	さいたま支社	E1004	ミルクココア	その他	1,300	60	78,000	
14	13	2018/1/9	柿本食品	さいたま支社	C1001	ドリップコーヒー	コーヒー	2,150	300	645,000	
15	14	2018/1/10	柿本食品	さいたま支社	C1002	カフェオーレ	コーヒー	1,700	300	510,000	
16	15	2018/1/10	ハウジング田中	前橋支社	T1001	煎茶	お茶	1,170	105	122,850	
17	16	2018/1/10	ハウジング田中	前橋支社	C1001	ドリップコーヒー	コーヒー	2,150	75	161,250	
18	17	2018/1/10	ハウジング田中	前橋支社	C1003	コーヒー無糖	コーヒー	2,000	75	150,000	

＞＞1-1-2　※該当するセルを囲み、上の①～③の引き出しを入れてください。

図 2 データクリーニングが必要な表の例

データクリーニングには、紙媒体のアンケートなど、手書きの内容をコンピュータ上でファイルに入力する作業も含まれます。その場合も、図 1 のようなレイアウトの表形式にまとめるのが基本です。

正しい分析のためには、回答の内容を正確に入力する必要がありますが、手書きの場合は、判読できない項目や記入漏れがあり得ます。このような項目は入力時にどうするのかといったルールを事前に決めておく必要があります。

また、コンピュータ上でいくつかのファイルに分けて入力されたデータを、一つのファイルに統合する作業もデータクリーニングに当たります。この場合、入力された表のフォーマットが異なることも多く、項目を統一するなどの修正作業が必要になることもあります。

(3)　Excel で集計や可視化などを行う

データクリーニングが終了したら、ようやく表のデータを元に集計や分析ができるようになります。

図 3 では、Excel のピボットテーブル機能を使って、各商品の売上金額を年別にクロス集計しています。データが正しい表形式に整備されていれば、このような集計表は 1 分もかからずに作成できます。

	A	B	C	D	E	F
1						
2						
3	合計 / 金額	列ラベル				
4		⊞2018年	⊞2019年	総計		
5						
6	行ラベル					
7	⊟お茶	26,478,900	29,420,550	55,899,450		
8	すこやか麦茶	3,240,000	3,442,500	6,682,500		
9	煎茶	9,090,900	10,021,050	19,111,950		
10	朝の紅茶	14,148,000	15,957,000	30,105,000		
11	⊟コーヒー	78,243,000	72,022,500	150,265,500		
12	カフェオーレ	17,068,000	16,957,500	34,025,500		
13	コーヒー無糖	34,300,000	31,200,000	65,500,000		
14	ドリップコーヒー	26,875,000	23,865,000	50,740,000		
15	⊟その他	36,984,300	41,408,400	78,392,700		
16	おいしい水α	16,186,800	17,933,400	34,120,200		
17	カップ麺セット	12,123,000	13,581,000	25,704,000		
18	ミルクココア	2,262,000	2,671,500	4,933,500		
19	熱々ポタージュ	6,412,500	7,222,500	13,635,000		
20	総計	141,706,200	142,851,450	284,557,650		

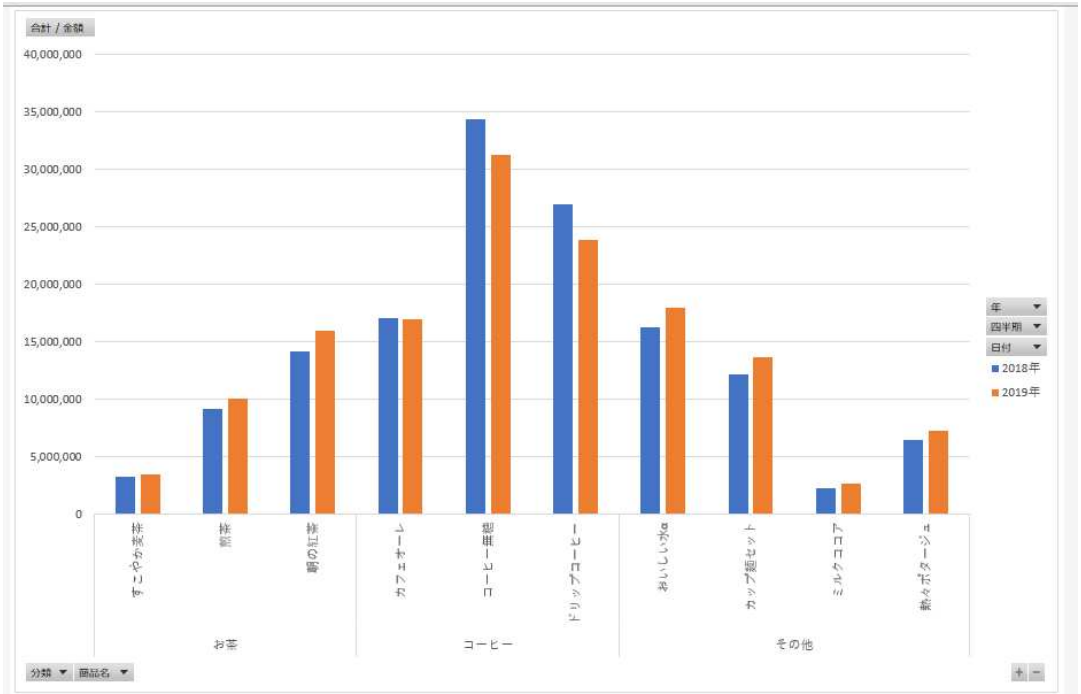
＞＞1-1-3

図３　ピボットテーブルで商品別に金額を集計した例

図 4 では、その内容をさらに棒グラフにして、金額の大きさを視覚的に比較できるようにしました。これを「可視化」といいます。

グラフでデータを可視化すると、どの商品がどのくらい売れているか、年による売上金額の違いがどの程度あるかといった内容を一目で把握できるようになります。

ここでは、「コーヒー無糖」、「ドリップコーヒー」という２つの商品の売上金額が群を抜いて大きく、2018 年、2019 年の両年ともに売上を牽引していることがわかります。



＞＞1-1-4

図４　縦棒グラフで商品別の金額を表示した例

このようにして、収集したデータをコンピュータでファイルとして保存し、データクリーニングを介してデータを整えれば、Excel の機能を使って分析を行うことができます。

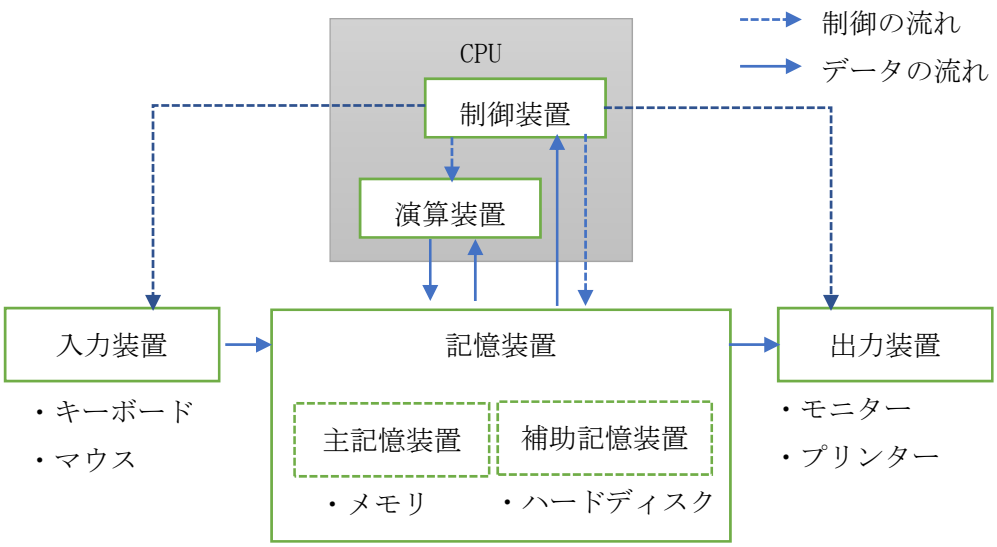
コンピュータは優れた分析ツールですが、データをコンピュータに取り込み、ファイルとして用意するだけでは不十分です。正しく分析できる状態でなければ、Excel などの BI（Business Intelligent）ツールとしてのソフトの機能を十二分に利用できない点に注意しましょう。

1.2. コンピュータの機能

観察や計測によって得られたデータの分析はコンピュータで行います。
ここでは、コンピュータの一般的な機能に加えて、コンピュータの内部で数値や文字等のデータがどのように処理されるかについて理解しましょう。

1.2.1. コンピュータの主要な装置

コンピュータのハードウェアは、一般に図のような５つの装置から構成されます。また、コンピュータ上で利用する主な機器の役割は以下のとおりです。



図●　コンピュータの５大装置と主な機器

(1)　CPU

CPU はコンピュータの心臓部で「中央処理装置」、「プロセッサ」とも呼ばれます。
データの処理を行う「演算装置」と、プログラムの命令に従って他の装置を制御する「制御装置」の二つの役割を持ちます。
CPU の性能は、コンピュータの性能の良し悪しに大きく影響します。CPU の処理速度を測る単位を「クロック周波数」といい、GHz(ギガヘルツ）という単位で表します。同じ構造の CPU ならクロック周波数が大きい方が処理速度は速く、高性能だと考えることができます。

(2)　メモリ

情報を記憶するためのスペースである「記憶装置」は、「主記憶装置」と「補助記憶装置」に分類されます。このうち、主記憶装置は、電源を切ると内容が消えてしまう揮発性の装置のことで、「メモリ」あるいは「メインメモリ」とも呼ばれます。
メモリは、制御装置や演算装置で利用されるデータを、一時的に格納しておく作業台のような働きをします。

(3)　ハードディスク

ハードディスクは、データを格納する倉庫の役割をするもので、磁気によって高速回転する円盤（ディスク）にデータを記録します。振動に弱く、一般にはコンピュータに内蔵されており、取り外しはできません。格納できるデータ容量は、数 GB(ギガバイト）～数 TB（テラバイト）が一般的です。
主記憶装置(メモリ）は揮発性のため、データの保存先には、電源を切っても内容が消えない不揮発性の装置を使います。これが「補助記憶装置」で、その代表格がハードディスクです。他の補助記憶装置については、「1.3.3　補助記憶装置の選び方」で紹介します。

(4)　キーボード、マウス

キーボードは、文字や数値、記号などのキーを使って、言葉や文章を入力したり、コンピュータを操作したりする装置です。マウスは左右のボタンに指を乗せて動かし、コンピュータへ指示を出すために操作する機器です。いずれも外部からコンピュータにデータを入力する「入力装置」に分類されます。

(5)　ディスプレイ

現在行っている操作の内容やパソコンの状態を画面に表示する装置です。コンピュータの処理の結果を外部に出す「出力装置」に分類されます。最近では、奥行きを取らない平面的なパネル型のディスプレイが多くなっています。

(6)　プリンター

コンピュータで作成した文書、表、画像などを印刷するための装置です。ディスプレイと同じく、「出力装置」の一つです。

家庭や職場では、ノズルからインクを噴射して印刷する「インクジェットプリンター」やコピー機のようにトナーを定着させて印刷する「レーザープリンター」が多く使われています。最近では、紙に書かれた文字やイラスト、写真などを読み取るスキャナーやコピー機能が付いた製品もあります。

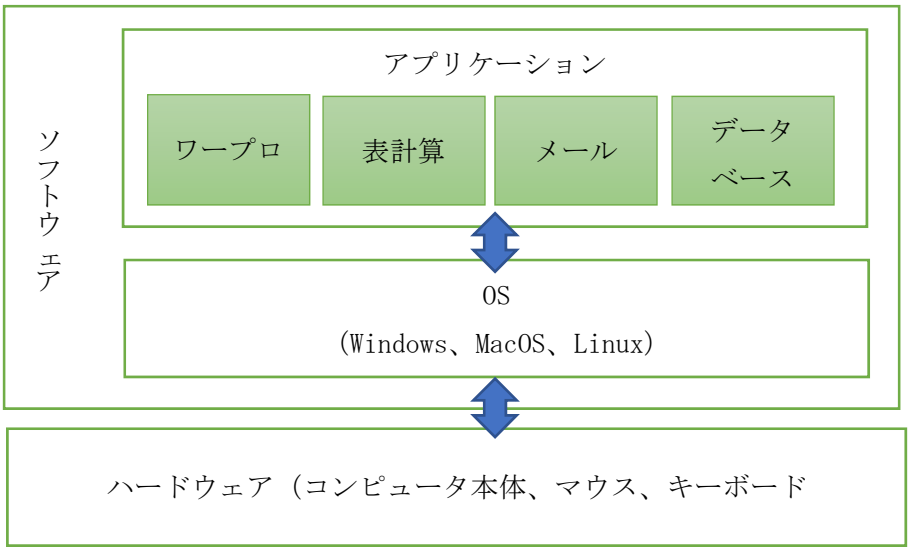
1.2.2. ハードウェアとソフトウェア

コンピュータは「ハードウェア」と「ソフトウェア」で構成されています。

「ハードウェア」とは、「1.2.1」で紹介したような装置や機械のことです。一方、「ソフトウェア」とは、コンピュータを動かすための動作手順を記述したプログラム(コンピュータプログラム）のことを指します。

コンピュータが仕事をするためには、ハードウェアが揃っているだけではなく、作業の指示を出すためのソフトウェアの存在が必要です。

ソフトウェアは、その役割から「OS(オペレーティングシステム)」と「アプリケーション」の 2 種類に分けることができます。



図●　ハードウェアとソフトウェアの関係

1.2.3. OS（オペレーティングシステム）の役割

「OS」は、「オペレーティングシステム」という言葉の略で、コンピュータ全般にかかわる基本的な機能をまとめたソフトウェアのことです。その内容から「基本ソフト」と呼ばれることもあります。

コンピュータの OS には、Windows や MacOS、Linux などの種類があり、パソコンの種類や用途に応じて利用します。コンピュータの起動や終了、ウィンドウやファイルの操作のしかたは、OS の種類により異なります。

OS には、公開された時期によってバージョンがあります。Windows のコンピュータでは、Windows7、Windows8.1、Windows10 といったバージョンがあり、バージョンによっても、OS の操作は多少異なります。

万一、OS に何らかのトラブルが発生して正常に機能しなくなると、コンピュータでは、通常どおりに処理ができなくなります。上の図●のように、OS はコンピュータ本体(ハードウェア）とソフトウェアをスムーズにつなぐための橋渡しをしていると考えましょう。

1.2.4. アプリケーション(アプリ）の役割

「アプリケーション」とは、コンピュータを利用してさまざまな作業を行うための個別のプログラムのことです。最近ではスマートフォンに倣って「アプリ」と省略されたり、「アプリケーションソフト」、「応用ソフト」などと呼ばれることもあります。

代表的なアプリケーションには、ワープロソフトの「Word」、表計算ソフトの「Excel」、データベースソフトの「Access」などがあります。そのほかにも、ゲーム、メールの送受信といった作業の内容に応じて用意された多種多様なソフトウェアの総称がアプリケーションです。これらのアプリケーションの中から、ユーザーは、自分が行いたい作業に適したものを選んで利用します。

アプリケーションと OS の関係は、2 階建ての建物によく似ています。基礎となる 1 階部分が OS で、2 階がアプリケーションです。2 階には、アプリケーションの種類の数だけ部屋があると考えるとわかりやすいでしょう。さらに、建物の下には、ユーザーが実際に操作を行うパソコン本体、マウスやキーボードといったハードウェアが存在します。

このように、OS とアプリケーションは互いに緊密に連携を取りながら仕事を進めるため、ユーザーは、コンピュータ上での作業をスムーズに行うことができるのです。

1.2.5. データを扱う際の留意点

ここでは、コンピュータの内部で、数値、文字、画像、音声、動画などのデータがどのように処理されるかを紹介します。コンピュータでデータを扱う際のそれぞれの留意点も併せて押さえておきましょう。

(1)　ビット、バイトとは

コンピュータ内部では、データを「0」と「1」からなる電気的な信号に対応付けて表現します。このため、数値、文字、画像、音声、動画などのデータもコンピュータ内部では、この０と１の組み合わせで表現されます。

なお、「0」か「1」のどちらかが表示されるこの最小単位を「１ビット(bit)」と呼びます。1ビットで表現される情報の量は０と１の２通りです。

また、8ビットをひとまとめにした単位を「１バイト (byte)」と呼びます。

コンピュータで、データの大きさなどを表現するには、ビットとバイトの２つの単位が使われます。

(2)　データサイズを表す補助的な単位

「バイト (B)」は、コンピュータ上でのデータサイズや記憶装置の容量などを表す単位として一般的に使われています。ただし、大きな数値を扱う際には、桁が多くなりすぎてわかりづらいため、下の表のような補助的な単位が使われることが一般的です。

補助単位を使うと、「24,000,000,000 バイト」を「24 ギガバイト」のように短く表すことができます。ビッグデータの活用などで、近年、データサイズはますます大容量化の傾向にあり、使われる補助単位も大きなものを目にする機会が増えています。

P64　表　データサイズの単位

単位	読み方	大きさ	備考
KB	キロバイト	10^3	1,000B に等しい
MB	メガバイト	10^6	1,000KB に等しい
GB	ギガバイト	10^9	1,000MB に等しい
TB	テラバイト	10^{12}	1,000GB に等しい
PB	ペタバイト	10^{15}	1,000TB に等しい

(3)　数値データの取り扱い

私たちが日常で使う数値は、10 を単位として桁が上がる 10 進数ですが、コンピュータ内部では、「0」と「1」だけで数値を表現できる 2 進数が使われています。したがって、コンピュータに入力された 10 進数の数値は、2 進数に置き換えてから計算が行われています。

2 進数では、2 を単位として桁が上がります。たとえば、10 進数の「2」は「2^1」なので 2 進数では「10」（イチゼロ）と表現されます。同様に、10 進数の「10」は「2^3+2^1」と考えます。これは、2 進数では「1000」＋「10」なので「1010」（イチゼロイチゼロ）になります。

ただし、2 進数の表示では桁が多くなりすぎるため、書きづらい・わかりづらいという難点があります。そこで、コンピュータでは、必要に応じて 16 進数を併用しています。

16 進数では 16 を単位として桁が上がり、「1、2、3、4、5、6、7、8、9、A、B、C、D、E、F」がそれぞれの桁に表示する「数字」として使われます。数字なのにアルファベットが混ざっているのは、アラビア数字の 1 桁の数値は 0 から 9 までしかないので、10 から 15 に当たる数値を A から F の英字で代用しているためです。

16 進数を使うと、2 進数の 4 桁を 1 桁で表現できます。たとえば、10 進数の「178」は、2 進数では「10110010」ですが、16 進数では「B2」となり、たった 2 桁で表すことができるのです。

このように、私たちが日ごろ使っている 10 進数の数値は、コンピュータの内部では 2 進数や 16 進数に置き換えられたうえで、計算や処理が行われることを知っておきましょう。

(4)　文字データの取り扱い

数値だけではなく、文字もコンピュータの内部では、「1」と「0」からなる 2 進数で表現されます。

文字を 2 進数で表すには、「A→1000001」、「B→1000010」のような文字と二進数の対応を決めておき、その対応表のルールに沿って文字を二進数に置き換えます。

この文字と二進数の対応の仕組みを「文字コード」と言います。現在、利用されている文字コードには、表のようなものがあります。

表●　主な文字コード体系

種類（読み方）	特徴
ASCII（アスキー）	米国規格協会によって定められたアルファベットと数字を表現するコード。7 ビットで文字を表現し、エラーチェック 1 ビットを含めて 1 文字を 8 ビットで利用する。
JIS（ジス）	JIS（日本工業規格）で規格化された文字コード体系。英数字、カタカナ、記号など半角文字を 8 ビットで、日本語のかなや漢字などの全角文字を 16 ビットで表現する。
Shift-JIS（シフトジス）	JIS を拡張した文字コード体系。16 ビットの漢字コードのデータと、8 ビットの英数字、カタカナなどのコードが混在している場合に識別がしやすくなった。日本国内のパソコン OS では、事実上の標準として利用されている。
EUC（拡張ユニックスコード）	UNIX という OS で日本語・朝鮮語・中国語などを扱うために制定された文字コード体系。日本語用のものを特に「EUC－JP」、「日本語 EUC」と呼ぶ。
Unicode（ユニコード）	一つのコード体系で各国語の文字を表現するために、ISO（国際標準規格）で標準化された文字コード体系。16 ビットでも文字数が足りず、一部の文字を 32 ビットに拡張するなど複雑化している。OS や Web での利用が広まっている。

文字データのトラブルに、意味不明の文字の羅列が画面に表示されてしまう「文字化け」があります。これは、Web サイトを閲覧している際や、受信したメール、インターネットからのダウンロードやメール添付で受け取ったファイルなどを開いた際に起こります。

文字化けの原因は、記述した際の文字コードと読む際の文字コードが一致していないことです。そのため、多くの場合は、記述した際の文字コードに合わせてデータを読み込みなおすと、正しく表示されます。

ブラウザやメールソフトの「エンコード」というメニューから「日本語（自動選択）」あるいは、「Unicode（UTF-8）」、「日本語（EUC）」、「日本語（シフト JIS）」といった主だった文字コードを順に選んでみるとよいでしょう。

なお、作成者のミスなどが原因であるため、情報を受け取った側では文字化けを直せない場合もあります。

(5)　画像データの取り扱い

画像データも、コンピュータの中では、「0」か「1」かで表現されます。デジカメで撮影した写真などのデータは「画素」と呼ばれる点の集まりになります。それぞれの点に色が付けられ、全体として見たときには写真の内容が表現される仕組みです。

画像における画素の数を「画素数」、「ピクセル」と呼び、画素数が多いほどより精密できれいな画像になります。ただし、情報量も増えるため、データサイズは大きくなります。

画像の画質とデータサイズは色の数によっても変わります。ディスプレイに表示される画像の色は、「RGB」と呼ばれ、赤（Red)、緑（Green)、青（Blue）の 3 色を混ぜて表現しています。このとき、１つの画素に何ビットの情報を持たせるかで色の数が決まります。割り当てるビット数を増やすと表現できる色の総数が増え、データサイズも増大します。

１つの画素に 24 ビットずつ割り当てたものを「フルカラー」あるいは「トゥルーカラー」と呼び、16,777,216 色を表現できます。これだけの色数があれば、人間の目に十分自然な画像を表現できるので、現在コンピュータやスマートフォン用の画像に多く使われています。

なお、画像には表のようなさまざまなファイル形式があり、どれを選ぶかによっても品質やファイルサイズが変わります。画像の用途に応じて最適なものを選ぶことが大切です。

表●　主な画像のファイル形式

ファイル形式	内容
BMP(ビットマップ)	Windows が標準でサポートしている画像の形式。圧縮していないためファイルサイズが大きくなる※
JPEG（ジェイペグ）	ISO（国際標準化機構）による国際基準の圧縮されたファイルフォーマット。多少の劣化はあるものの圧縮率が高く、デジカメ写真の保存用フォーマットとして一般に広く使われている。
GIF（ジフ）	圧縮されたファイルフォーマット。256 色以下の画像を扱うことができる。色数の少ない図、イラストなどの画像に向く。
PNG（ピング）	圧縮されたファイルフォーマット。でファイルサイズが小さい。画像の劣化が少なく、透過処理もできるため、Web サイトで図やイラストなどに使われることが多い。

注※　ファイルサイズを小さくする方法については、「1.6　ファイルの圧縮解凍」を参照。

(6)　音声データ、動画データの取り扱い

音声データは、音波という一連の波であり、アナログな情報です。音声をコンピュータの内部でデジタルデータとして扱うには、一定の間隔で値を測定して取り出す「標本化(サンプリング)」や、測定値を数値化する「量子化」という処理を行います。

動画データは、一定間隔で変化する画像と音声の組み合わせで構成されます。動画の中の1枚1枚の画像を「フレーム」と呼び、1秒あたりのコマ数を「フレームレート」と呼びます。フレームレートが高いほど、動画の動きは滑らかになります。

音声や動画のデータは、データ量が膨大であるため、ファイルサイズが極端に大きくなります。これらのデータをやり取りする際には、データの圧縮が不可欠です。これについては、「1.6　ファイルの圧縮/解凍」で紹介します。

1.3. ファイル管理

では、あなたが保存したデータは、コンピュータ上でどのように管理されるのでしょうか。ここでは、データ管理に欠かせないフォルダーやファイルの扱い方について理解しましょう。

1.3.1. ファイルとは

アプリケーションで作成した数値、文字、画像、音声、動画等のデータは、すべて「ファイル」という単位で、コンピュータの記憶装置上に保存されます。

ファイルのアイコンは、データの種類に応じて絵柄が異なるので、どのアプリケーションで作成したファイルなのか、どういう種類のファイルなのかを見た目で判別できます。

図○　ファイルアイコンの例

＞＞画像 1-3-1

ファイルの名前は、正確には「ファイル名」と「拡張子」で構成されます。

たとえば、上の図の左端のファイルは、Word で作成した文書ファイルなので、「報告書.docx」というファイル名になります。前半の「報告書」が、データを保存するときに自分で指定できる名前の部分です。

一方、その後ろに続く「.docx」のような、「.（ピリオド)」と英数字の部分を「拡張子」と呼びます。「拡張子」は、下の表のように、ファイルの種類によってあらかじめ決まっていて、ファイルの保存時に自動的に指定されます。

拡張子	ファイルの種類	対応するアプリケーション
.docx	Word 文書ファイル	Word
.xlsx	Excel ブック	Excel
.accdb	Access データベース	Access
.jpeg　.jpg	画像ファイル	ペイントなどの画像処理ソフト

表○　主な拡張子と対応するアプリケーション

拡張子は、ファイルの種類をコンピュータが識別するために必要なもので、拡張子の種類によって、アイコンの絵柄が決まります。

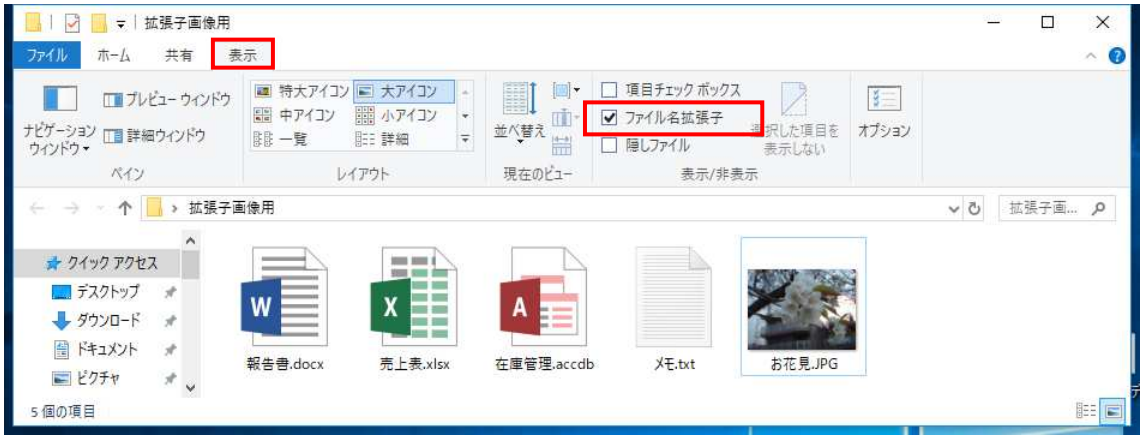
また、ファイルアイコンをダブルクリックしたときに、どのアプリケーションでファイルを開くのか、その組み合わせも拡張子ごとに決められています。これをファイルの「関連付け」と言います。

ファイル名を変更する場合などに、うっかり拡張子を削除してしまうと、この関連付けが破壊されて、ファイルを開けなくなるといった不具合が発生することがあります。このようなトラブルを予防するため、拡張子は通常では表示されない設定になっています。

拡張子を確認したい場合は、次の手順で、ファイル名に拡張子を表示することができます。

<手順>

① いずれかのウィンドウで「表示」タブをクリックし、「ファイル名拡張子」にチェックを入れる。



>>画像 1-3-2

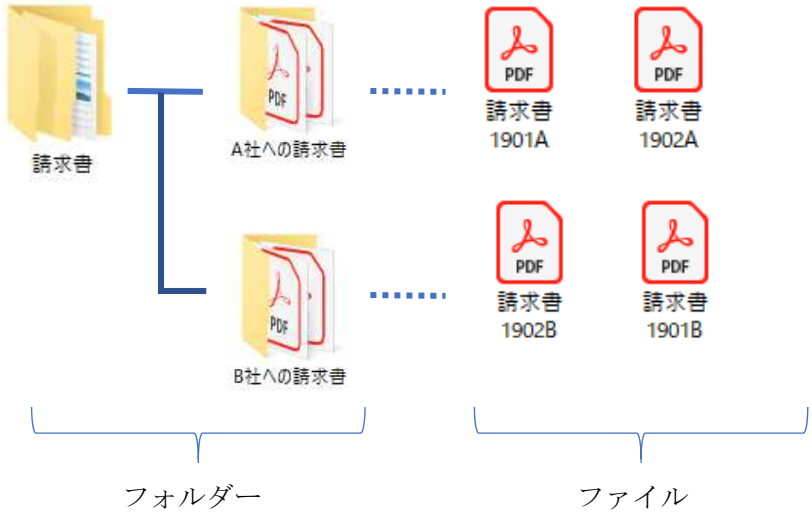
これで、上の図のように、そのコンピュータ上では、ファイル名の後ろに拡張子が表示されます。

1.3.2.　　フォルダーとは

ファイルの数が増えると、目的のファイルを探すのが大変になります。そこで、コンピュータでは、ファイルを入れ物に入れて分類できるようになっています。この分類のための入れ物が「フォルダー」です。ちょうど私たちが物を箱に入れて整理整頓するのと同じです。

フォルダーは必要に応じて自由に作成できます。また、フォルダーの中にフォルダーを作って、ファイルを階層的に管理することも可能です。

下の図では、請求書のファイルを探しやすく管理するために、得意先別に「A 社への請求書」、「B 社への請求書」といったフォルダーを作りました。さらにその上位に「請求書」というフォルダーを作り、「A 社への請求書」、「B 社への請求書」などのフォルダーをその下に配置しています。これは、大きな箱の中に小さな箱を入れて中身を分類するのと同じ構造です。



図○　階層になったフォルダーで請求書のファイルを管理する例

＞＞1-3-3-1～1-3-3-4 で作図ください。

1.3.3.　　ドライブとは

階層構造で管理されたフォルダーやファイルは、最終的にはハードディスク、USB メモリ、CD、DVD といった記憶装置に保存されます。このファイルやフォルダーの最終的な保存先となる記憶装置のことを「ドライブ」と呼びます。

使用しているコンピュータのドライブは「エクスプローラ」で確認できます。

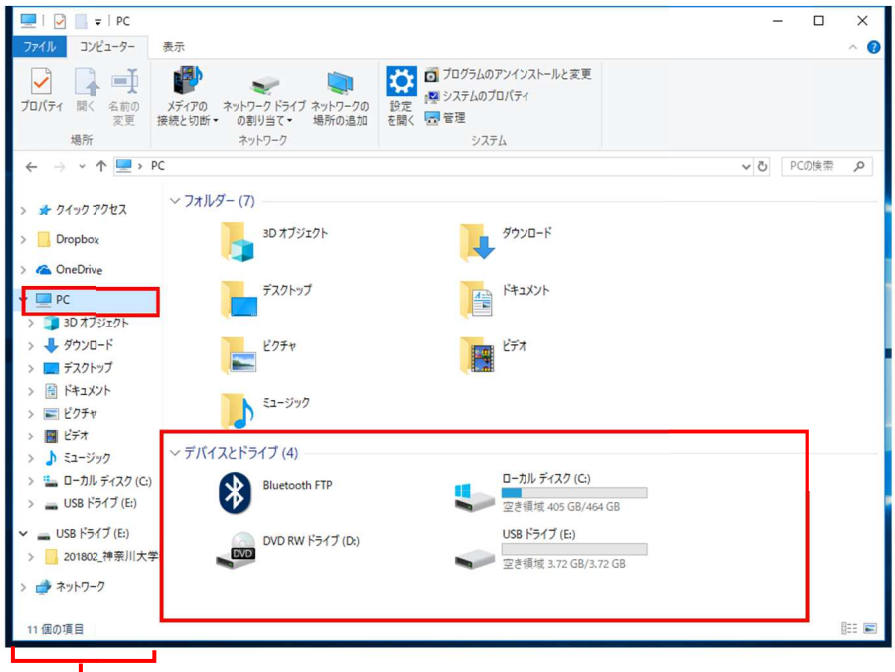
あなたが使用しているコンピュータのドライブを確認してみましょう。

＜操作＞

① タスクバーにピン留めされたアイコン「エクスプローラ」をクリックし、「エクスプローラ」を起動します。

＞＞1-3-3-5

② エクスプローラの「ナビゲーションウィンドウ」で「PC」をクリックすると、「デバイスとドライブ」に、現在接続している機器やドライブが表示されます。なお、ドライブの構成内容は、コンピュータの状態によって異なります。



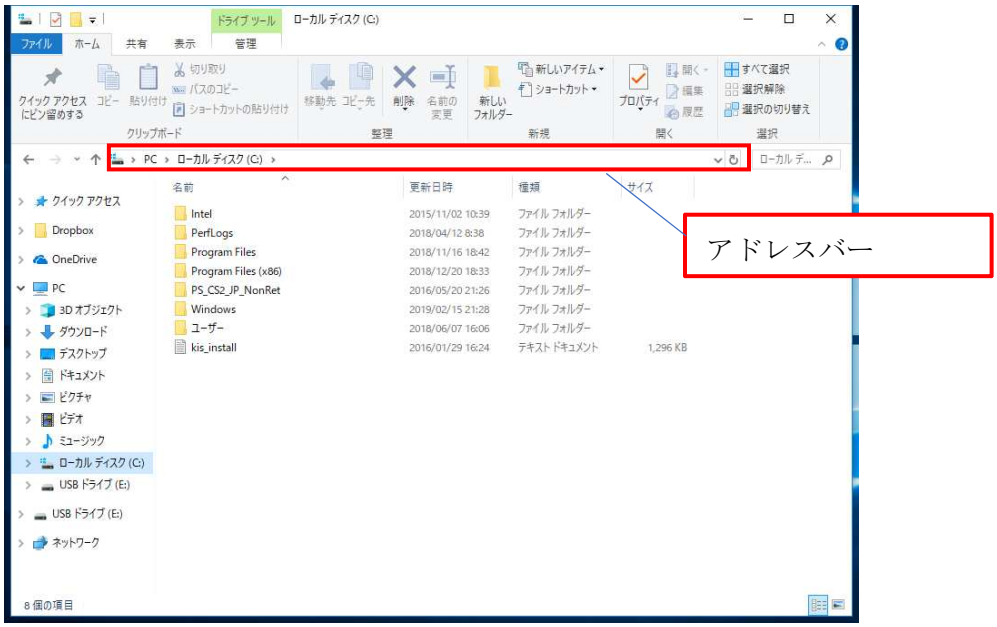
ナビゲーションウィンドウ

＞＞1-3-3-6

ドライブの名称は「C」、「D」といったアルファベット 1 文字と「：(コロン)」を付けて表示されます。下の図では「ローカルディスク（C:)」はコンピュータ内蔵のハードディスクを指しています。また、「USB ドライブ（E：)」はコンピュータに接続した USB メモリを意味します。記憶装置の種類は、アイコンの絵柄で判別できます。

＞＞1-3-3-6

ドライブの内容を見たいときは、ドライブのアイコンをダブルクリックします。
下の図は、「ローカルディスク（C：)」のアイコンをダブルクリックした結果です。



＞＞1-3-3-7

エクスプローラに C ドライブ（ハードディスク）の内容が表示されます。
なお、ウィンドウ上部の「アドレスバー」には、ドライブやフォルダーの階層が表示されます。ここを見ると、現在表示されているフォルダーの階層の位置を確認できます。また、アドレスバーで階層をクリックすると、その階層のフォルダーやファイルがウィンドウに表示されるので、階層間の移動が楽にできます。

1.3.4.　補助記憶装置の選び方

コンピュータの主記憶装置は、揮発性のメモリであるため、電源を切ると内容が消えてしまいます。そこで、電源を切っても内容を残しておくことができるよう、揮発性ではないメモリを補助的に利用します。これを「補助記憶装置」といいます。

主な補助記憶装置には、表のようなものがあります。データを保存する際には、用途やデータの容量などに応じて適切に補助記憶装置を選択しましょう。

種類	補助記憶装置の名称	データ容量	容量別単価	読み書き速度	携帯性
磁気ディスク	ハードディスク(内蔵型)	◎	○	○	×
	ハードディスク(外付け型)	◎	○	○	△
フラッシュメモリ	SSD(内蔵型)	○	×	◎	×
	USB メモリ	○	×	◎	○
	メモリカード(SD)	○	×	◎	◎
光学ディスク	CD(CD-R、CD-RW)	×	◎	×	◎
	DVD(DVD-RW)	△	△	×	◎
	Blu-ray Disc(BD-RE)	○	○	△	◎

表　主な補助記憶装置の比較

(7)　磁気ディスク

磁気により円盤状のディスクにデータを記憶させる記録メディアが「磁気ディスク」です。代表例であるハードディスク (HDD)はコンピュータに内蔵され、取り外しができないものが一般的ですが、外付け型の製品もあります。ただし、外付け型であっても大きいため、携帯性にはやや劣ります。

読み書きスピードが速く、記憶容量は数 GB～数 TB までと幅広い製品があります。内蔵型のものは、コンピュータにおける日常でのデータ保存に使われるほか、OS やアプリケーションのプログラムをあらかじめ格納しておき、コンピュータの部品の一部として稼働するのが一般的です。外付けの製品は容量の割に低価格であることから、コンピュータ全体のバックアップなど大容量データの保存に多く使われます。

(8)　フラッシュメモリ

記憶領域に半導体を用いたメモリのことを「フラッシュメモリ」といいます。読み書きスピードが高速ですが、価格は高めの傾向にあります。代表例である USB メモリは、コンパクトで携帯性に優れ、現在、データの持ち運び用メディアとして多く使われています。紛失時の情報漏洩を予防するため、セキュリティロック機能が付いたものなど、バリ

エーションに富んだ製品が販売されています。容量は数GBから１TB程度までのものがあり、予算に応じて使い分けることができます。

「SD カード」と総称されることの多いメモリカードには、容量によってさらに複数の種類に分かれます。容量は数 MG から数 GB 程度で、携帯電話やデジタルカメラの画像、動画などを保存する媒体として広く普及しています。

フラッシュメモリを用いた内蔵型記憶装置に「ソリッド・ステート・ドライブ（SSD)」があります。HDD よりも高速で読み書きができる、動作音が静かであることから昨今、コンピュータやスマートフォンで利用されるようになりました。ただし、ディスク容量がハードディスクに比べると小さく、価格が高いことがネックです。

(9)　光学ディスク

光の反射を用いてデータの保存を行う記憶メディアが「光学ディスク」です。CD、DVD、Blu-ray Disc などの種類があり、これらはデータ記録の規格が異なるため保存容量などに違いがあります。

CD のデータ容量は 650MB、700MB が一般的で、一度だけ書き込みができる CD-R と何度でも書き込み可能な CD-RW があります。容量は小さめですが、ディスク 1 枚当たりの価格が格段に安く携帯性もよいため、資料データなどの配布用に使われることが多いメディアです。

DVD は、CD よりも容量が大きく、片面 4.7GB、両面 9.4GB のものが一般に普及しています。その記憶容量から、動画の保存、映像の記録に用いられます。また、Blu-ray Disc は新世代の光ディスク規格と言われ、25GB～100GB とさらに大容量のデータを記録できます。

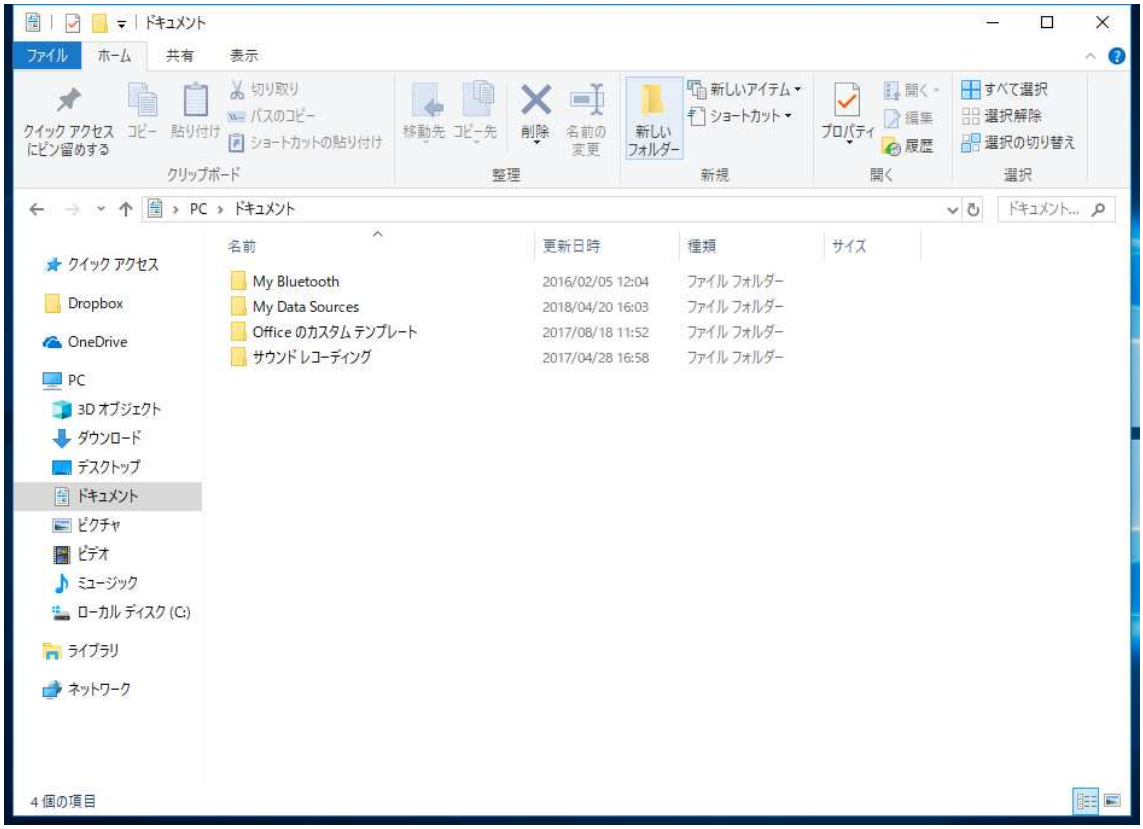
1.3.5.　　ファイルやフォルダーの操作

ファイルをわかりやすく管理するためには、フォルダーを作成し、ファイルを適切に分類して保存することが大切です。よく使うファイルやフォルダーの操作を知っておきましょう。

(10)　フォルダーの新規作成

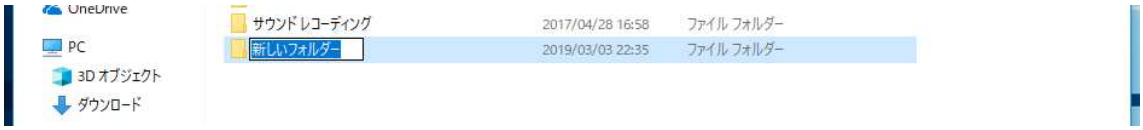
フォルダーは必要に応じて自由に作成できます。ここでは、ドキュメントに「請求書」という名前のフォルダーを作成します。

<操作>



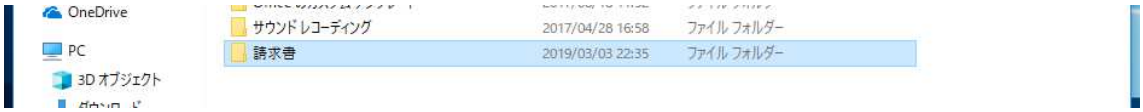
>>1-3-5-1

① エクスプローラのナビゲーションウィンドウで「ドキュメント」をクリックして開き、「ホーム」タブの「新規」グループにある「新しいフォルダー」をクリックします。



>>1-3-5-2

② 「新しいフォルダー」というアイコンが表示されたら、フォルダー名を「請求書」と入力し、Enter キーを押します。



＞＞1-3-5-3

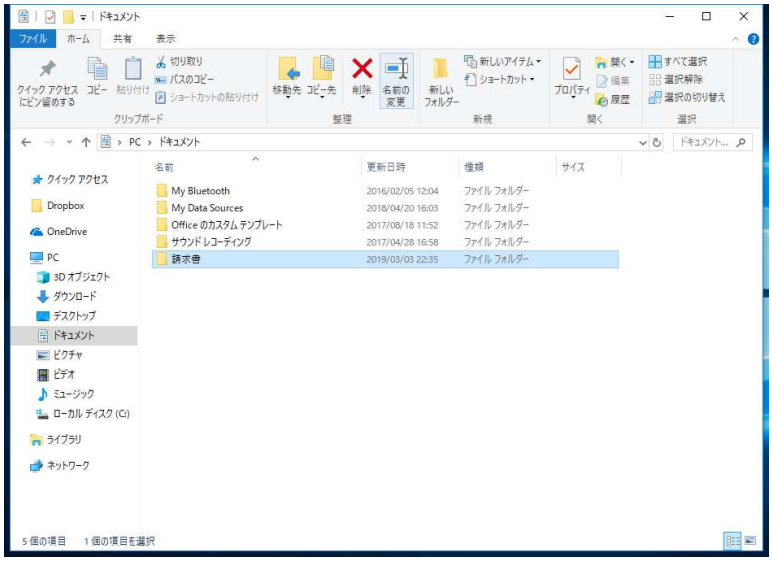
③ 「請求書」という名前のフォルダーが作成されました。

＜コラム＞ ファイル名やフォルダー名には、次の半角の記号は利用できません。使わないように注意しましょう。 ¥（円記号）　＊（アスタリスク） /（スラッシュ）　？（疑問符） ：（コロン）　＜　＞（不等号） ！（感嘆符）

(11) ファイル、フォルダーの名前の変更

フォルダー名は後から変更できます。ここでは、「請求書」から「書類」へフォルダー名を変更しましょう。なお、同じ手順でファイル名も変更できます。

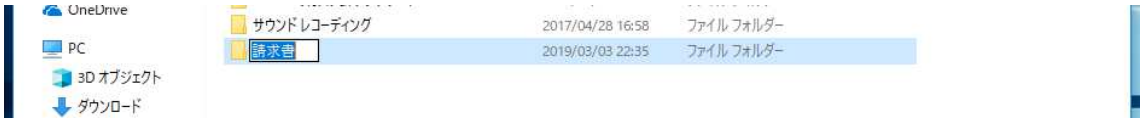
＜操作＞



＞＞1-3-5-4

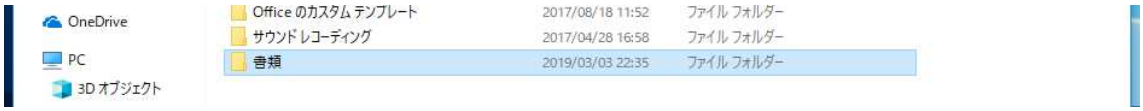
① 「請求書」フォルダーを選択し、「ホーム」タブの「整理」グループから「名前の変

更」をクリックします。



＞＞1-3-5-5

② 現在のフォルダー名が反転表示になります。キーボードから新しいフォルダー名を「情報」と入力し、Enter キーを押します。



＞＞1-3-5-6

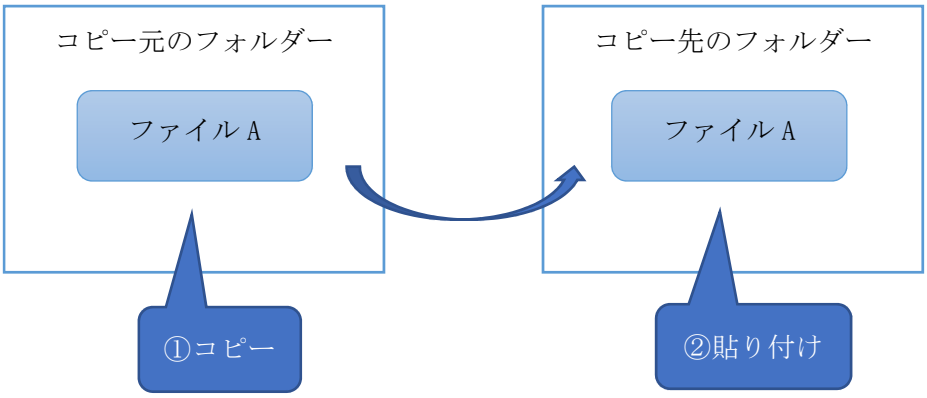
③ フォルダー名が「情報」に変更されました。

＜コラム＞ 変更したいフォルダー名を 2 回クリックするか、F2 キーを押してもフォルダー名を変更できます。

(12) ファイル・フォルダーのコピー

コンピュータの不具合や紛失などに備えてファイルやフォルダーをコピーしておくことを「バックアップを取る」といいます。大切なデータは、日頃からまめにコピーをしてバックアップを取るように心がけましょう。

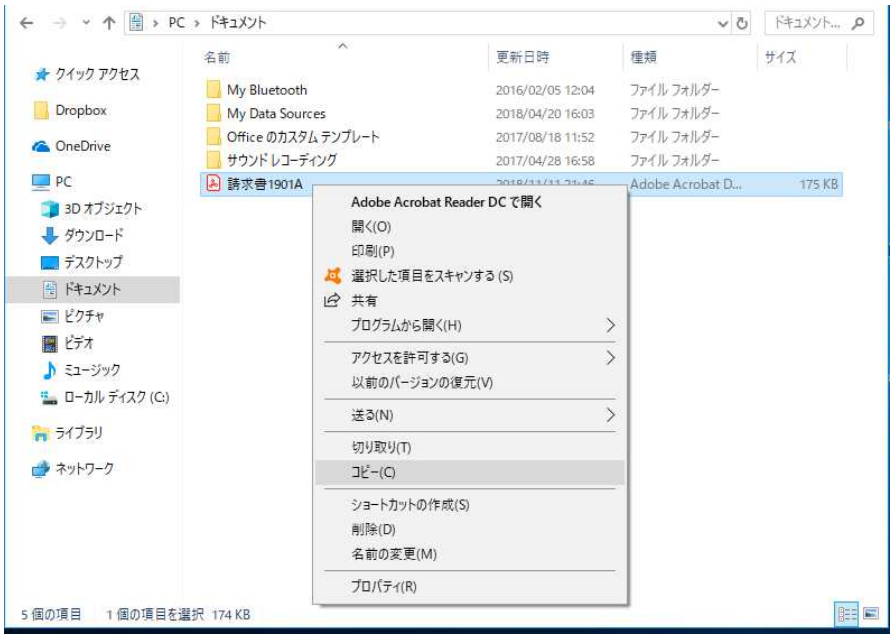
ファイルやフォルダーをコピーするには、「コピー」と「貼り付け」の操作をします。複製を作りたい対象を「コピー」して、コピー先の場所で「貼り付け」を実行します。



ファイル A の複製ができる

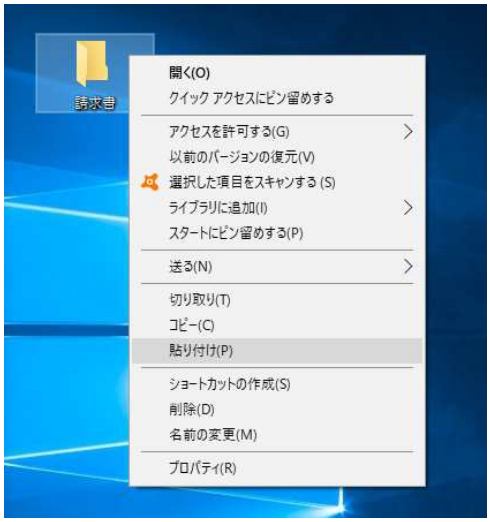
ここでは、ドキュメントにある pdf ファイル「請求書 1901A」をデスクトップのフォルダー「請求書」にコピーします。

＜操作＞



＞1-3-5-7

① コピーしたいファイルを右クリックし、ショートカットメニューから「コピー」を選択します。



＞1-3-5-8

② コピー先であるフォルダー「請求書」を右クリックし、ショートカットメニューから「貼り付け」を選択します。

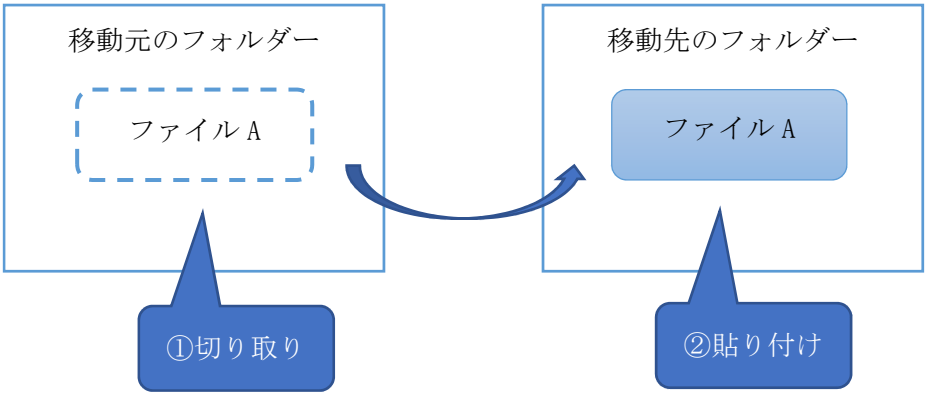
＞1-3-5-9

③ 「請求書」フォルダーを開くと、pdf ファイル「請求書 1901A」がコピーされています。

なお、フォルダーをコピーした場合は、そのフォルダーに保存された下位のフォルダーやファイルも合わせてコピーされます。

(13) ファイル・フォルダーの移動

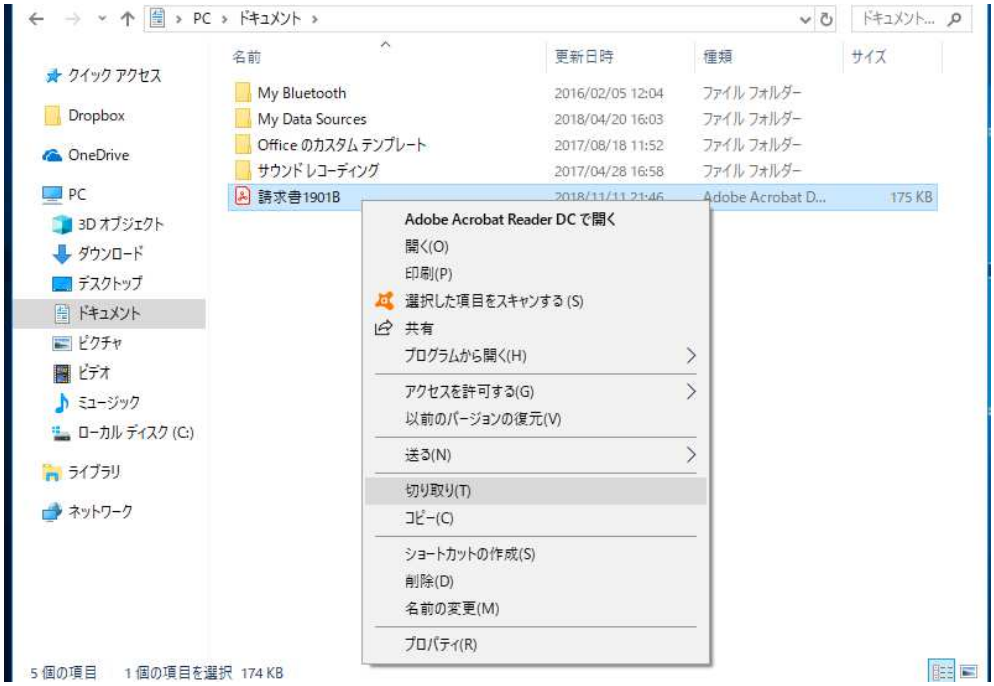
ファイルやフォルダーを別の場所に移動するには、「切り取り」と「貼り付け」の操作をします。移動したい対象を選んで「切り取り」を実行し、移動先の場所で「貼り付け」を実行します。移動の操作が終わると、対象のファイルやフォルダーは元の場所からなくなります。



ファイル A は移動元フォルダーからなくなる

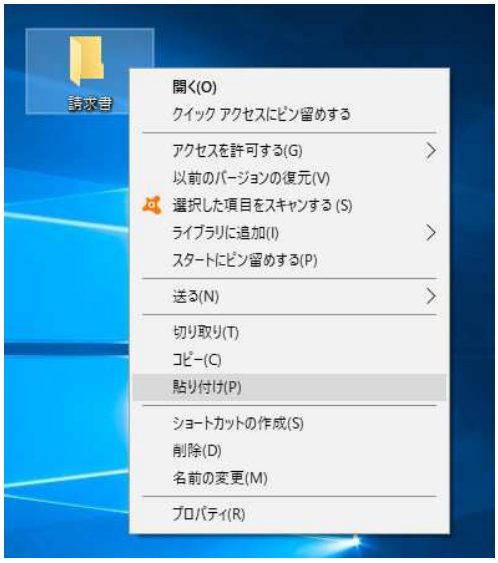
ここでは、ドキュメントにある pdf ファイル「請求書 1901B」をデスクトップのフォルダー「請求書」に移動します。

＜操作＞



＞1-3-5-10

① 移動したいファイルを右クリックし、ショートカットメニューから「切り取り」を選択します。



＞1-3-5-11

② 移動先であるフォルダー「請求書」を右クリックし、ショートカットメニューから「貼り付け」を選択します。これでドキュメントから pdf ファイル「請求書 1901B」がなくなり、「請求書」フォルダーに移動します。

＜ショートカットキー＞
・コピー　→　Ctrl＋C
・切り取り　→　Ctrl＋X
・貼り付け　→　Ctrl＋V

(14)　ファイル・フォルダーの削除

コンピュータでは、不要になったファイルやフォルダーは、安全のため 2 段階の操作で完全に削除される仕組みになっています。

まず、不要になったファイルやフォルダーを削除するには、対象となるフォルダーやファイルを選択して「Delete」キーを押します。こうして削除された対象は、いったんデスクトップの「ごみ箱」に移動します。

「ごみ箱」にあるファイルやフォルダーは、元の場所に復元することができます。ごみ箱からさらに対象を削除すると、完全になくなり、復旧することはできなくなります。

ただし、ごみ箱に入るのは、ハードディスクドライブのファイルだけです。USB メモリなど外付けの補助記憶装置に保存されたファイルやフォルダーは、「Delete」キーを押すと、ごみ箱には入らずに、即座に内容が削除されます。これらのデータは復旧することができないので注意しましょう。

ここでは、デスクトップにある「請求書」フォルダーを削除し、その後復旧します。

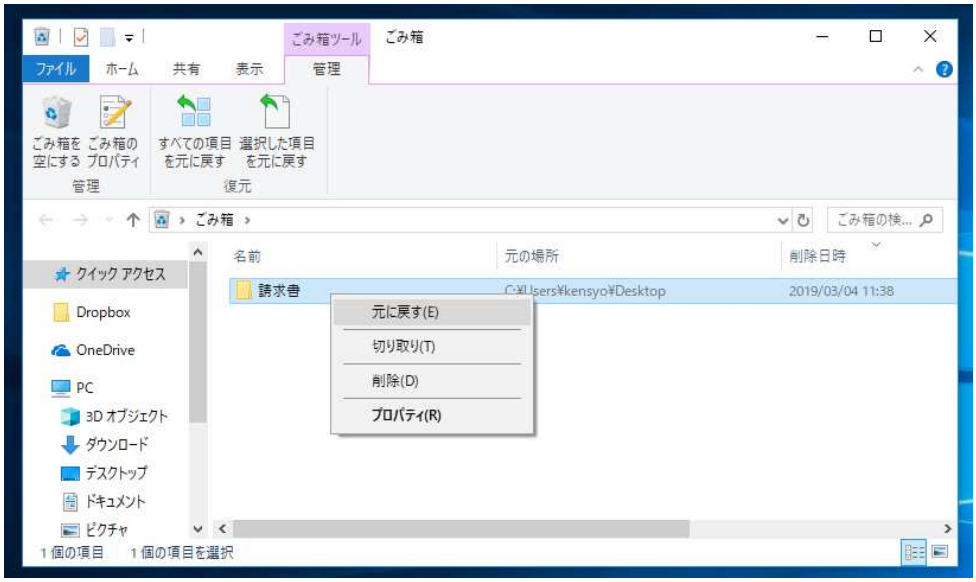
＜操作＞

＞1-3-5-13

①「請求書」フォルダーを選び、「Delete」キーを押します。あるいは、「請求書」フォルダーを「ごみ箱」までドラッグします。

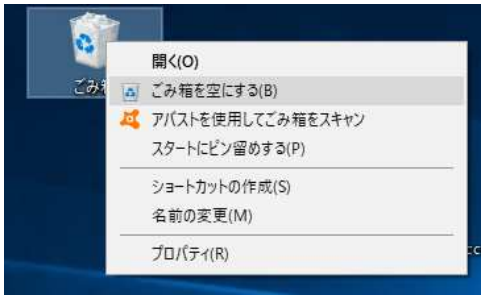
＞1-3-5-14

②「ごみ箱」に削除したデータが移動します。この段階では、「請求書」フォルダーを復旧できます。



＞1-3-5-15

③「請求書」フォルダーを復旧するには、ごみ箱を開き、「請求書」を右クリックして、ショートカットメニューから「元の場所に戻す」を選択します。これで「請求書」フォルダーがデスクトップに復旧されます。



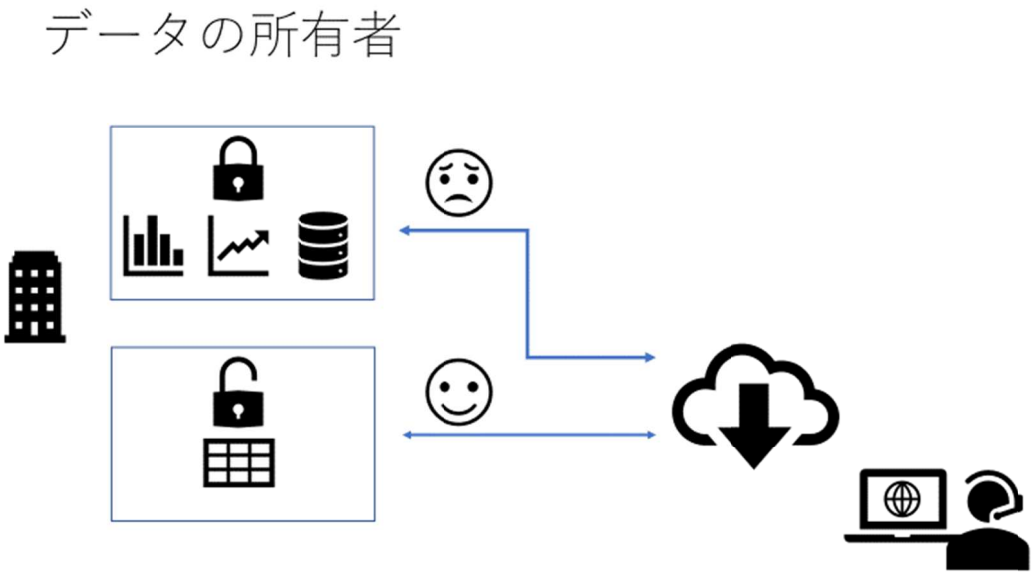
＞1-3-5-17

「ごみ箱」の内容を完全に削除するには、ごみ箱を右クリックし、ショートカットメニューから「ごみ箱を空にする」を選択します。完全に削除するかどうかを尋ねるメッセージが表示されるので、「はい」をクリックすると、ごみ箱の内容が完全に削除されます。

1.4. ブラウザの利用

ここでは、様々なデータを収集するうえで留意すべき点と、効率よくデータを収集するために欠かすことのできないブラウザの活用方法、またインターネット経由によるデータ収集と活用における注意点について学びます。

1.4.1. データの所有者



さて、この章に入る前に、1.1.1. 「データ、情報、ファイルとは何か」の項で学んだデータの種類について復習しておきましょう。

同項では、様々な手法（手書き、インタビュー、システム登録、インターネットなど）で蓄積されたデータが、様々な形態（電子データ、報告書、書籍等）で存在することを学びました。そして、その「データの所有者」の存在についても触れました。

全てのデータには原則的に「所有者」が存在し、その所有者がデータに閲覧権限や利用範囲を設定しています。

例えば、POS システムにおいて集計されたデータは、システムを所有する店舗、または店舗をチェーン展開している企業の所有物であり、その企業の許諾無しに我々が目にすることはできません。

仮に何らかのシステムの不具合や、元従業員でシステムにアクセスできる ID やパス

ワードを使って見ることができてしまった場合でも、その時点で閲覧権限を持たない人が、企業の許諾無しに情報資産へアクセスした場合、犯罪に問われます。

不正アクセス禁止法

http://elaws.e-gov.go.jp/search/elawsSearch/elaws_search/lsg0500/detail?lawId=411AC0000000128

これについては、●●●のところで詳しく解説します。

また従業員であって、データへのアクセス権限を有する人でも、個人情報や営業情報など、流出した際に企業の経営に重大な影響を与えるデータの取り扱いには十分な配慮と漏えい防止に努めなければなりません。

1.4.2　公開データ

では、我々が閲覧できるデータにはどんなものがあるでしょうか？

皆さんは「国勢調査」という言葉を聞いたことがありますか？この「国勢調査」とは“国内の人口・世帯の実態を把握し、各種行政施策その他の基礎資料を得ること”を目的とするもので、総務省統計局の管轄で5年毎に行われるアンケート調査です。

この「国勢調査」では国の行政施策を立案するために必要なデータを収集する目的で、アンケート調査時において本邦内（日本国内）に常住している者を対象に実施さ

れます。国も企業同様に、施策を立案する際には詳細な、かつ膨大なデータを活用しています。

（図 1-1）総務省統計局ホームページ　「平成 27 年国勢調査」

https://www.stat.go.jp/data/kokusei/2015/index.html

この「国税調査」では、「世帯員に対する事項」として氏名、男女の別、出生の年月をはじめとする 13 項目と、「世帯に対する事項」として、世帯の種類や世帯員の数などの 4 項目について記入式、またインターネットを使ったアンケートによって調査します。このデータは集計後、インターネットを利用した公開に続き、報告書として刊行されています。（※但し、個人を特定する情報は含まれていません）

（演習 1）

「国勢調査」のデータを使うと、どんな仮説が導き出せるのか３つ考えてみましょう。

データの種類	これがわかる	こういうことが考えられる

一例を挙げれば、【出生の年月】のデータにより、その地域の平均年齢等がわかり、公共設備投資の施策内訳（子供向けの公園が必要か？高齢者向けの医療機関が必要か？）などが変わってきます。また【国籍】によって、外国籍の方が多いエリアの行政としての施策（どの言語で行政サービスをより充実していく必要があるか？）なども変わってきます。

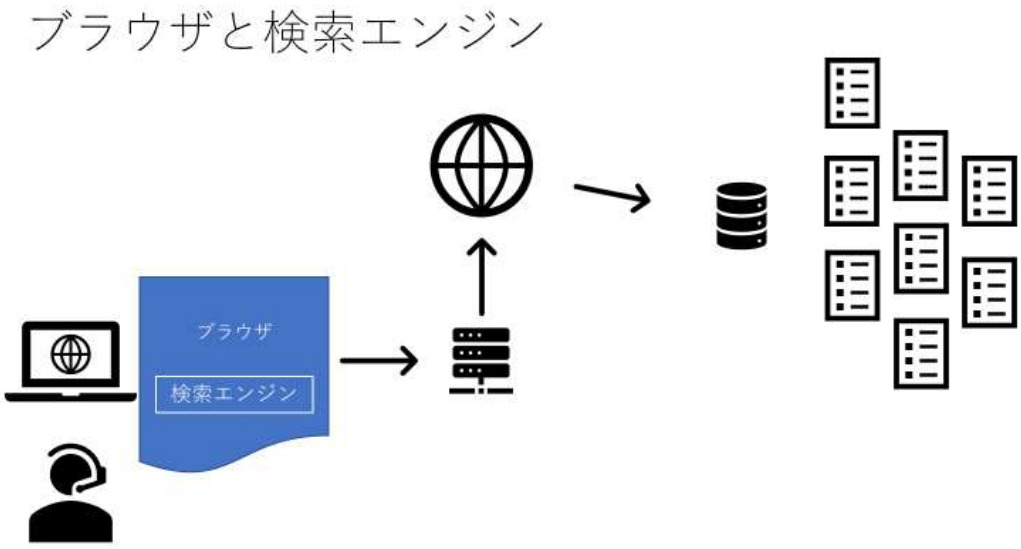
また、このようなデータ以外にも、企業や団体から公開されている様々なデータが無数に存在しています。

一例を挙げると、株式市場に上場している企業は、金融商品取引法に則り決算報告書の開示が義務付けられ、金融庁が運営する EDINET

（http://disclosure.edinet-fsa.go.jp/）にて誰でも閲覧できるようになっています。
国内外の金融機関や個人投資家は、この決算報告書や決算短信といったデータをもとに、企業の業績や可能性、またリスクを読み取り、投資の判断に活用しています。
このような公開データは近年、インターネット上で公開されていることが多く、そのデータを収集するために有効な手法がブラウザ経由で入手する方法です。

1.4.3.　ブラウザと検索エンジン

「Google 先生に聞いてみる」なんて言葉があるくらい、皆さんは日々当たり前のようにインターネットを使って調べ物をしていると思います。
でも、「Google 先生はどうやって知りたいことを調べているの？」という質問に正確に答えることはなかなかむずかしいのではないでしょうか。ここではデータ分析を行う上で、最低限知っておきたい「検索の仕組み」について学びます。



（図１－２）ブラウザと検索エンジン

まず、皆さんが「インターネット」という言葉から連想するコンピュータの画面は、「ブラウザ」と呼ばれるアプリケーションソフトです。代表的なブラウザとしてはGoogle、Yahoo!、Bing 等で、このブラウザがパソコンにインストールされていないと、Web ページの閲覧ができません。皆さんがお持ちのスマホには予めインストールされているため、普段あまり意識することがありません。
このブラウザには「検索窓」と呼ばれる入力スペースがあり、探したい情報のキーワードを入力し、実行ボタン（虫メガネのアイコン）をクリックすると、システ

ム側で検索エンジンが働き、インターネット経由で様々なデータをもとに該当すると思われる Web ページの候補をリスト（ヘッダ情報）表示してくれます。

（図 1-3）リスト（ヘッダ情報）表示例

1.4.4.　表示順位

検索キーワードに基づいて検索結果のリストが表示されましたが、そもそもこのリストはどのような順番で並んでいるのでしょうか？

一般的に、人は「検索結果の表示画面の 1 ページ目の中にある情報をクリックするか、2 ページ目まで見て、それらしきページ（ヘッダ情報）が見つからない場合は、別のキーワードで再検索を行う」と言われているほど、検索結果の 1 ページ目に表示されることはそのページを使って様々な営業活動を行っている企業にとって非常に重要

なことです。

図 1-3 を見ていただいて分かる通り、4 番目までのページには【広告】と表示されています。これは企業や団体が「データ分析」というキーワードで検索された際に、自社のページが上位に表示されるよう、広告費を支払っていることを意味します。一概に広告中心の Web ページには有益な情報が無いとは言えませんが、最終的な目標が「顧客獲得」であることから推測されるに、データ分析に必要な公開データの収集にはあまり適さないページであると思われます。

また広告費を支払わないまでも、上位表示されるためのテクニック（SEO 対策）を施し、少しでも上位表示されるための努力をしている企業が多いのも事実です。

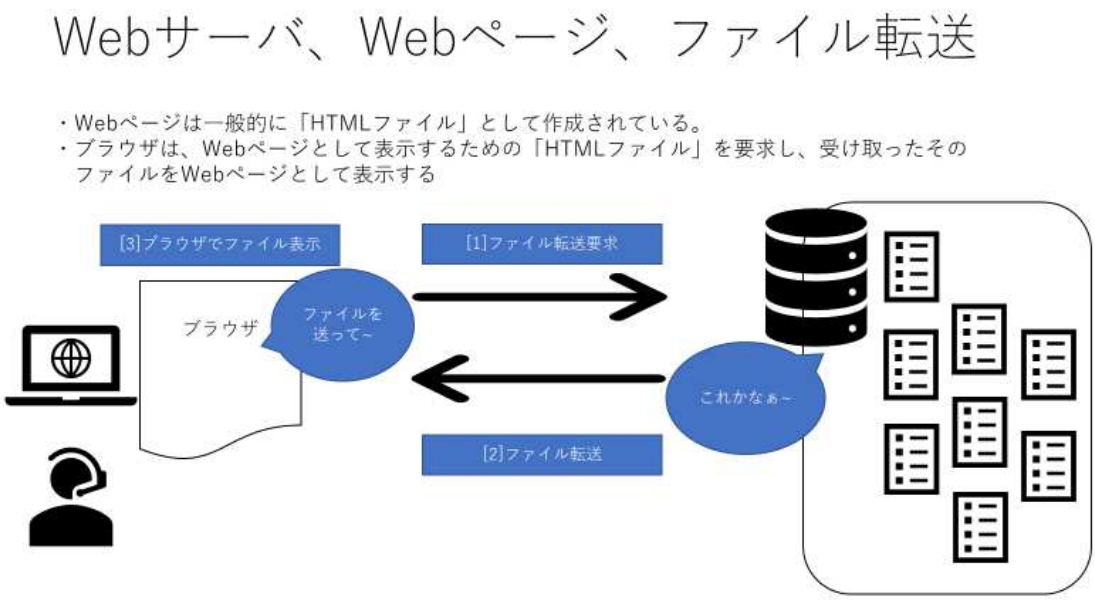
余りにも広告ページが多い、またはお目当てのページでは無さそうであれば、2 ページ目、3 ページ目を調べてみるか、検索キーワードを複数入力する等して、お目当てのページを見つけるようにしましょう。

なお、リストに表示された検索結果は、クリックしても良いページだけ表示されているとは限りません。フィッシング詐欺など、インターネットを利用した詐欺行為に巻き込まれる恐れもあります。これについては、【1.4.7. サイトアクセス時の危険性について】のところで詳しく解説します。

1.4.5　Web ページの表示

検索してリスト（ヘッダ情報）表示された中から、「これかな？」と思われるリストをクリックすると、皆さんがいつも見る Web ページが表示されます。

この動きをシステム的に説明すると、以下の図のようになります。



（図 1-4）Web サーバ、Web ページ、ファイル転送

この際、検索キーワードをもとに「どのページを選択して表示するか」を司っている機能が、クローラと呼ばれる各ブラウザが持っているプログラムで、Web 上にある様々なデータ（文書、画像、音声等）を周期的に収集し、そのブラウザが有するデータベースに登録していきます。

代表的なクローラとしては Googlebot、Bingbot などがあり、１兆を超えると言われる Web ページをデータベースにせっせと登録し、検索結果としてユーザーに戻してくれています。

「どの検索結果を表示するか」については、各ブラウザによって異なります。これは各ブラウザのクローラが持つアルゴリズムが異なるためで、同じ検索キーワードで検索しても表示順位が異なることがあります。

学校の PC で使用しているブラウザだけでなく、自宅の PC やスマホでの検索も行ってみることをお勧めします。

（図 1-5）Bing による検索結果表示画面（「データ分析」）

1.4.6.　データ収集方法

では、いよいよデータを収集する効率的な手法について学んでいきましょう。一口に「効率的」と言っても、何をもって効率的と呼ぶのかを定義しないと始まりません。

また効率とともに、収集したデータの信ぴょう性、データの引用方法など、分析を行う上で留意すべきポイントについて正しい理解をしておかないと、正しい結果が得られないばかりか、「迷惑をかける分析結果」になりかねません。

本書では、以下の項目を提示し、その項目に沿って解説を進めて行きます。

1）検索手順（キーワードの作り方）
2）検索環境（検索作業の環境構築）
3）データの信ぴょう性
4）データの入手方法
5）データの引用

1）検索手順（キーワードの作り方）

ブラウザによる検索において、「キーワードとして何を選ぶか？」が検索結果を大きく変えます。「そのページってどうやって見つけたの？」と、友人の検索結果に驚いたり、探しているページがなかなか見つからずイライラしたりした経験はありませんか？これはほぼ「キーワードの作り方」の差と言えます。

（演習２）

ではここで、実際にブラウザで検索を行い、以下の条件に適合する Web ページを探してみましょう。自分が「コレだ！」と思うページの URL を以下に書き留めてください。皆さんの検索が終わったら、実際のそのページを比較してみてください。

【テーマ】

・2018 年に、どの国の外国人が日本に多く来日したか？上位 10 カ国を調べたい。

※ただし、ランキングの根拠やそのページを公開している企業名等が明らかなものに限ります。

※検索結果で表示された Web ページに掲載されているリンクから辿っても OK です。

検索日：

検索キーワード（検索窓に入力したまま記入ください）：

URL：

ページの名前：

データ名：

データの提供元：

付加情報（ランキング以外の有益な情報、詳細な情報があれば記入ください。：

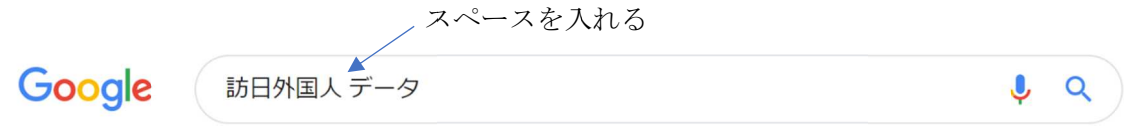
さて、みなさんはどんなページを見つけることができましたか？隣の人と同じページでしょうか？それとも別のページでしょうか？

検索キーワードの作り方ひとつで結果が変わるとすれば、「どうやって効率よく

データを収集」は「どのように検索キーワードを設定するか」が重要になってきます。
ここでは一般な検索手順（キーワードの作り方）を学んでいきましょう。

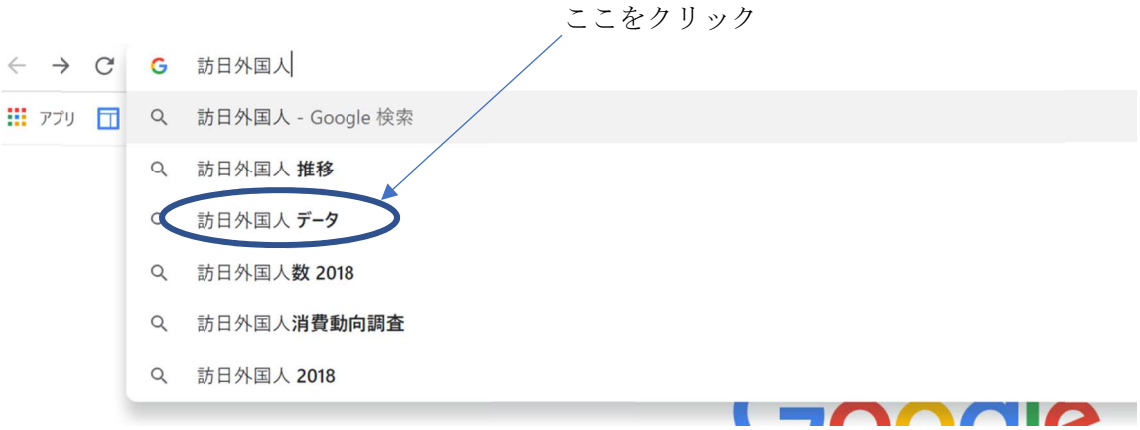
(1)一般的な検索方法

AND 検索とは、「あるキーワードと、もう一つ以上のキーワードを複数入力して検索を行う」方法です。最近は AND の代わりにスペースを入れて検索する方法でも同様の結果が得られます。なお、より精度を高めるためにはキーワードを 3 つ、4 つと増やすことも有用ですが、あまりに多いと該当ページ数が減り、逆にお目当てのページが表示されない可能性もありますので注意しましょう。



OR 検索とは「どちらか」または「いずれか」のキーワードで検索をかける方法です。
前述の AND 検索が複数のキーワードをすべて含むページを検索するのに比べ、OR 検索では「どちらか」、「いずれか」のキーワードを含むページを検索するため、表示候補数としては AND 検索よりも多い傾向です。AND 検索を行った際にお目当てのページが表示されない場合や、より範囲を広げて再検索を行う際に有用です。

また検索窓にキーワードを一つ入れると、AND 検索の表示候補が複数表示される場合があります。これは、そのキーワードで検索されたことが多いことを示しているものなので、その中に自分が考えているキーワードがあればそちらをクリックして検索するのも有効な手段です。



(2)検索結果を絞り込みたいときの検索方法

Web ページには様々な情報が混在していることが一般的です。アジアの事について調べたいのに、ヨーロッパやアフリカの情報まで含んでいるページも多く、検索結果を絞らないと調べたいデータを見つけるまでに膨大なページを探さなければなりません。

このような時に役立つ検索方法として、「完全一致検索」や「NOT 検索」があります。ある程度絞り込むができれば検索作業の効率化が図れます。ここでは多くの候補ページからの絞り込みに重点を置き、「NOT 検索」について紹介します。

NOT 検索は、検索キーワードでヒットする Web ページの中から、外したいキーワードを使って絞り込みする方法です。例えば、横浜駅から羽田空港までのアクセス方法で、電車以外の方法を検索するときには、「横浜駅から羽田空港 アクセス -電車」と入力することで、バスや車といった電車以外でのアクセスが上位表示され、お目当てのページにアクセスしやすくなります。

なお、分析で使用する機会の多い統計資料データは、お目当ての項目以外のデータも含めてまとめられている場合が多いため、NOT 検索を使用しても効果が得られないこともあります。

(3)検索キーワードの作り方

AND検索やOR検索の手法がわかっても、キーワードが的確でないとお目当てのデータになかなかたどり着けないことがあります。ではどうやってキーワードを見つければ良いのでしょうか？

（演習2）において、「2018年に、どの国の外国人が日本に多く来日したか？上位10カ国を調べたい」というテーマで検索していただきましたが、その際には「訪日外国人」という言葉を選びました。ちなみに似たような意味を持つ言葉を使って検索した際の、候補ページ数を見てみましょう。

訪日外国人　　約　29,900,000件
来日外国人　　約 235,000,000件
入国外国人　　約　 9,480,000件
外国人旅行客　約　46,300,000件
（※2019年2月末時点、Googleにて検索）

検索キーワードによって候補ページ数が 2 桁も違ってくるというのを理解いただけると思います。

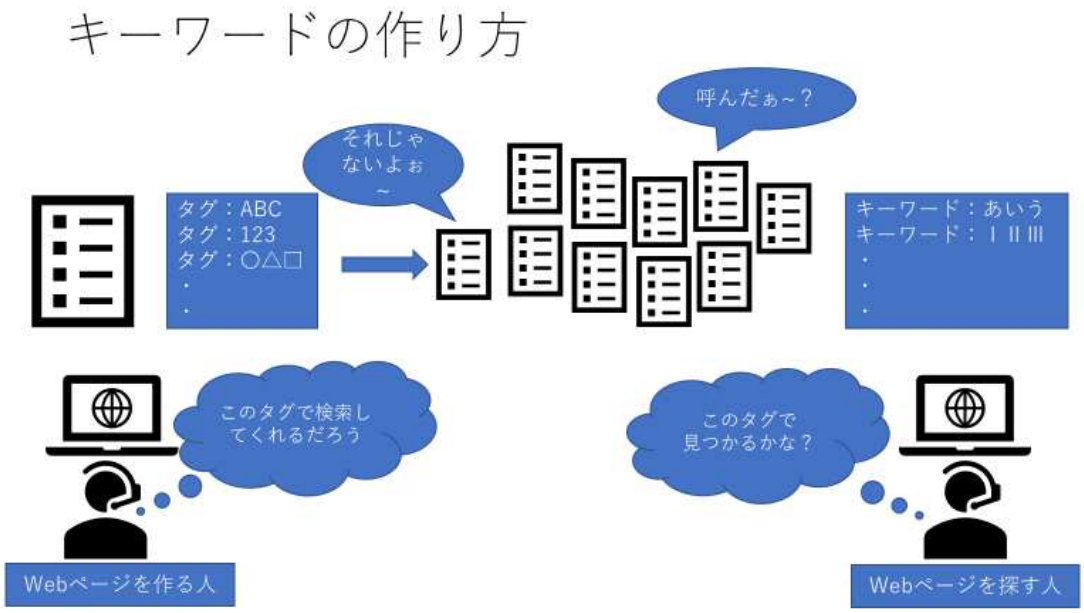
では本題の検索キーワードの作り方に入って行きましょう。
検索キーワードを考える上で大切なポイントがいくつかありますが、ここでは

・一般的によく使われている言葉かどうか
・探したいデータが持っていると思われる特徴的な言葉かどうか
・検索エンジンが推測しやすい言葉かどうか

を解説したいと思います。

Web ページには、その情報を必要としている人が検索する際にヒットされやすい言葉言葉（タグ）を埋め込んでいます。どんなに有益な情報を Web ページとして掲載したとしても、見てもらえないのであればインターネット上にある「ただのゴミ」としてそこにあるだけだからです。

そのタグを埋め込む人は、様々なタグを考え、Web ページに埋め込んでくれていますが、そのほとんどが一般的によく使われている言葉、またはそのページを必要としてくれる人が使っている言葉が中心になります。



（図 1-6）キーワードを選べないと出会えない

ブラウザを活用してデータ収集を行う際には、この一般的な言葉、そして使われて言葉葉を作る必要がありますが、この言葉は普段、皆さんが各メディア（テレビ、新聞、インターネット等）で目にするものであり、日頃からキーワードとして使える言葉かどうかを意識していないと、いざという時に頭に浮かんできません。いわば、あなたの語彙力が試されていると思ってください。

（演習 3）

以下のデータを探すために、検索キーワードとして考えられる言葉を書き出し、検索結果から回答を導き出しましょう。

問題１．（初級）

「愛媛県松山市内で、車で遊びに行ける公園はいくつある？」

（※施設名称に「公園」が付いていなくても、「公園」と認知されている場所）

キーワード	
解答	箇所

さて、答えがわかりましたか？ではもう少し難易度を上げて行きます。

問題２．（中級）

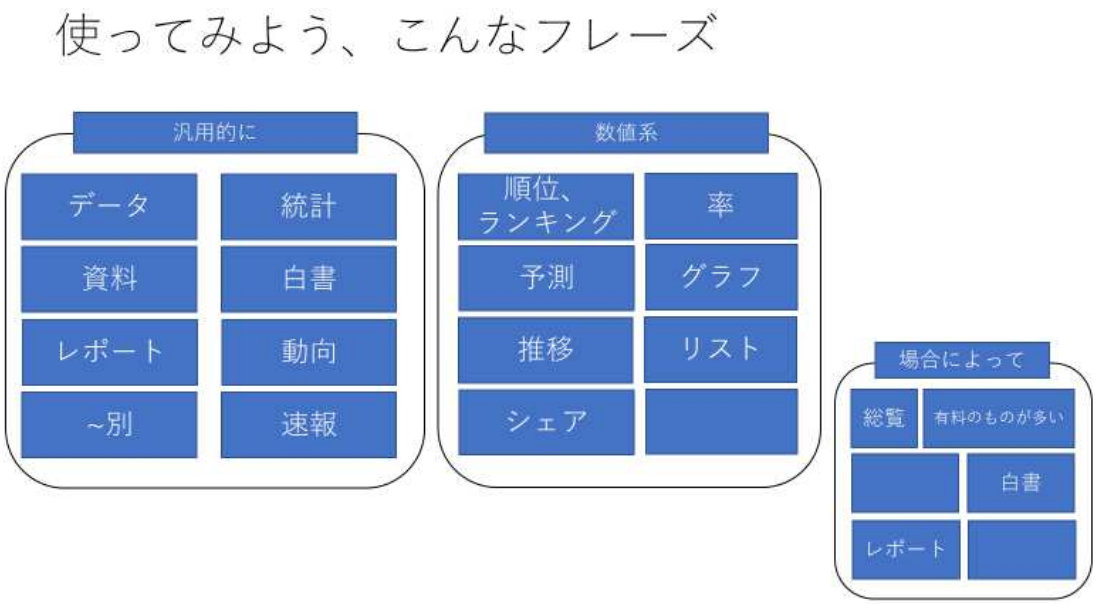
「高知県にお住いのご家庭が持っている車の平均台数は？」

キーワード	
解答	台

さぁ、どうでしたか？わかりましたか？

検索して表示されたページには、皆さんにとって「聞いたことがある言葉」も「初めて見た言葉」もあると思いますが、いくつものページを見ることで、「このデータはこんな言葉を使ったら出てくるかな？」と、ドンドン試して検索する経験を積んでください。

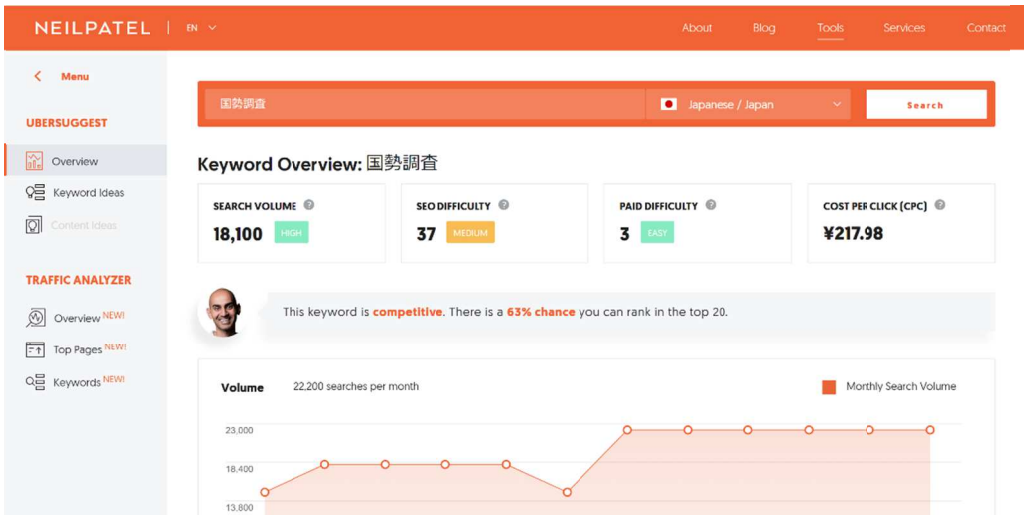
データ検索する際のキーワード候補として、以下の文字を活用してみましょう。



また、「どんなキーワードが検索に用いられているのか」にも興味が湧くと思いますが、残念ながらブラウザは教えてくれません。(※有料で提供してもらうことは可能な場合もあります)

そんな時は無料で教えてくれる無料サービスを活用してみましょう。

NEIL PATEL（https://neilpatel.com/）というサービスがあり、ここでは自分が検索しようする候補のキーワードを入力すると、そのキーワードで検索された月次の件数等と、類似の検索で使われたキーワードとその回数が確認できます。一つの参考にしてみても良いでしょう。



２）検索環境（検索作業の環境構築）

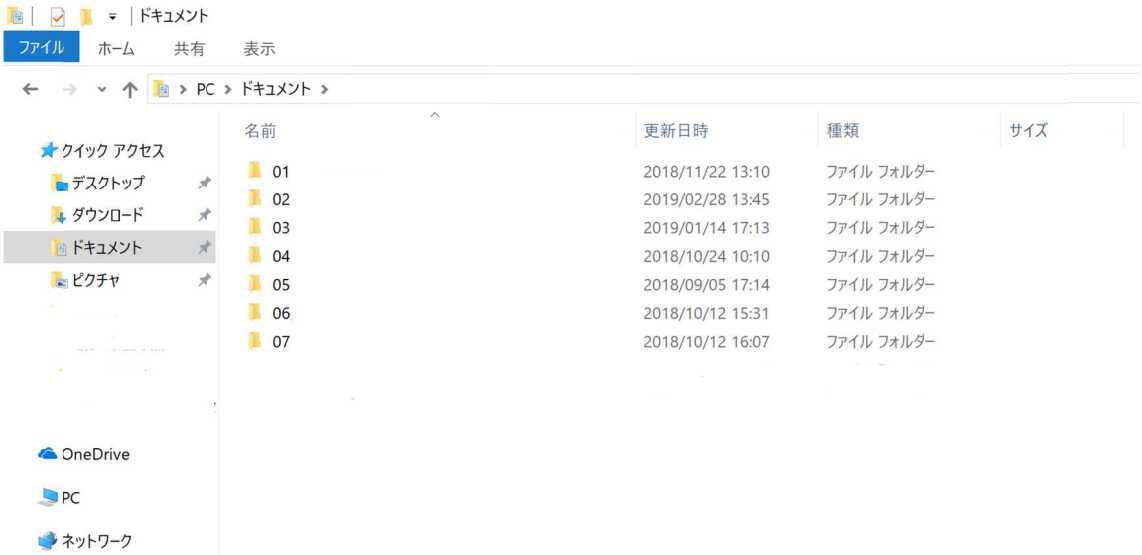
「ブラウザを活用したデータ収集」は端末（PC、スマホ等）を使用して行いますが、効率よく作業が行えるための「環境構築」も重要です。一人で行う作業、また複数で行う作業など、作業状況を問わず確認しておくべき「環境構築」について学びましょう。

（※ここでは主に PC をベースにした解説を行いますが、スマホでも同様に構築が可能です）

(1)デスクトップはきれいにしよう

ブラウザで検索した情報を、Word やメモ帳、また Excel にコピペしたものをデスクトップに置きっぱなしにすると、気がつけば置き場がない・・・なんてことになりかねません。

デスクトップはいわば「机の上」。作業場である机の上が書類でゴチャゴチャしていると作業の効率も上がりませんね。デスクトップはあくまで「作業台」として考え、様々なファイルは「ドキュメント」の中に作業ごとのフォルダを作り集約しておくことをお勧めします。



（図 1-7） ドキュメント内のフォルダ

集約した後、改めてファイルを開こうとしたときに「どのファイルだったかなぁ・・・」とファイルが迷子になってしまうことがあります。ファイルに名前を付ける際には、規則性を持たせ、探しやすくすることも大切です。

(2)検索結果のページは保存しておく

検索して探し出したページは、何回も見直すことも多いはずです。そんな時、「あれぇ？

あのページってどこだったっけ？」とならないように、検索結果の Web ページをフォルダを作成して保存しておきましょう。

ブラウザに標準で用意されている「ブックマークバー」や「お気に入り」に登録することができますが、ブラウザを活用してデータ分析をおこなう方は検索しに行くページも多いはず。そんな時は「ブックマークバー」や「お気に入り」にそのまま置くのではなく、検索結果ページごとにフォルダ管理してしまいましょう。

関連ページを一つのフォルダにまとめておくことで、何回も検索する必要がなくなり、効率的な検索作業が行えます。

Google を例に紹介いたします。ブラウザの検索スペース右端にある★をクリックすると、上記ウィンドウが開きます。ここで[その他]をクリックすると新しいフォルダが作成できます。

ここで[新しいフォルダ]をクリックし、検索したページを集約するフォルダを作ります。

アクティブになっている（青く反転）ところに、フォルダの名前を入力し、[保存]をクリックしたら保存完了です。保存する前に Web ページにつけられている名前が長すぎたりする場合は、[名前]の枠にある名前を修正しても良いでしょう。

この機能を使うことで、苦労して探し出したページが保存、整理でき、同じページを開く際にもアクセスが容易になります。

3)データの信ぴょう性

やっと見つけた Web ページですが、さて、そのページに掲載されているデータを鵜のみにする前に、一旦考えてみましょう。それはそのデータの「信ぴょう性」です。データ分析を行う上で、必要なポイントがいくつかありますので、それを見て行きましょう。

(1)出所が確かか？

これは〇〇〇で詳しく説明しますが、そのデータを Web ページに掲載している企業、団体はそのデータとどのような関わりを持っているのかによって、「信ぴょう性」が異なります。例え同じ名前で掲載されているデータでも、掲載している出所が確かなことが必要です。

近年、個人でも簡単に立派なページを作れるようになり、あたかも公式サイト的な見え方をするページもあります。また自社、自分の Web ページへのアクセスを増やし、アフィ

リエイトを活用して収益を上げている企業や個人の方もとても多く、「見つけたぁ！」と思って飛びつくのは危険です。こんな順位でページを眺めてみましょう

信ぴょう性が高いと言われるデータ出所

◇国や都道府県の機関、もしくはそれに準ずる組織
- 省庁（例）内閣府、総務省、経済産業省、他
- 外郭団体（例）独立行政法人、特殊法人、特別の法律により設立される民間法人
- 財団法人、社団法人
- 都道府県が設置している協会

◇公共性の高い組織
- 組合、連合会、基金

◇その分野で著名な企業

次に、信ぴょう性が疑われるデータについて考えてみましょう。

そのデータ、信ぴょう性はありますか？

- Webページにおける、情報発信者が明確でないもの
  →「誰が発信した情報か」が特定できない

- データの収集方法が不明なもの
  →集計データの場合、「収集方法」「時期」「対象」などの説明が明記されていないもの

- 明らかに「剽窃（ひょうせつ）」された文章データ
  →Webページ自体の信ぴょう性が低い

まず、検索したページに掲載されている情報について、「誰が発信した情報か」が特定できないデータは信ぴょう性が低いと言わざるを得ません。Twitter や個人ブログでハンドルネームを使用している場合も、一部のケース（発信者が特定でき、信用するデータと第三者が認識できるもの）を除いて同様に扱われてしまいますので注意しましょう。

次に、集計データ等の収集方法が不明なものです。公的機関が公表している集計データ

には、データの収集方法についての説明や注釈が記載されています。主に「収集方法」「時期」「対象」などが明記されていて、どの程度の信ぴょう性を保持しているのかが判断しやすいデータです。

（表：総務省統計局 国勢調査の概要説明ページ）

一方､｢高校生 100 人が選んだベスト〇〇〇｣といったデータについて考えてみましょう。あなたはこのデータの、何が「問題になりそうだ」のと思いますか？そもそもどんな収集方法で集めたデータかについて説明が書かれていたとしても、高校生という対象の大きさに比較して、たった 100 人のデータで作られたランキングに対して、あなたはどの程度の信ぴょう性を感じるでしょうか？ウソではないが、データとしての信ぴょう性に疑問を持たれてしまうデータは分析には不向きと考えるべきです。

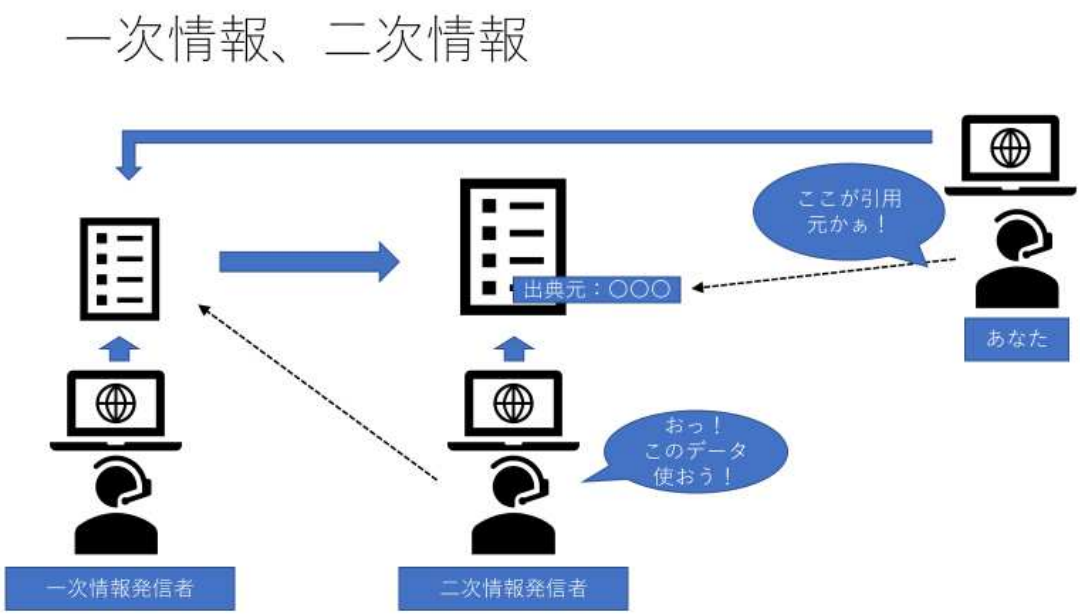
また、用いているデータの出所ははっきりしたとしても、Wikipedia 等からのコピペが多い等、そのページ自体の信ぴょう性が低い場合も少なくありません。すぐ見てわかるページばかりではありませんが、企業や団体よりも個人のページに多くみられる傾向ですので、念のため留意しておきましょう。

（※一部の団体は、その団体の事業性格上、あえて注釈が無くても信ぴょう性が担保されているページもあります）その収集方法に収集者の思想等の偏りがないかどうかが判断の一つです。

		乗用車				貨物車			
		普通	小型	計	本年累計	普通	小型	計	本年累計
ダイハツ	合計	13	4,181	4,194	8,611	0	0	0	0
	前年比	56.5	159.4	158.5	183.3				
日野	合計	0	0	0	0	4,699	1,178	5,877	10,453
	前年比					109.7	118.4	111.4	108.4
ホンダ	合計	16,907	18,427	35,334	65,431	0	0	0	0
	前年比	108.4	90.4	98.2	102.1				
いすゞ	合計	0	0	0	0	5,103	1,407	6,510	11,377
	前年比					104.7	135.3	110.1	107.1
レクサス	合計	5,479	0	5,479	11,334	0	0	0	0
	前年比	88.8		88.8	89.8				
マツダ	合計	11,598	3,653	15,251	27,377	55	1,176	1,231	2,176
	前年比	139.6	69.2	112.2	102.6	101.9	111.3	110.8	106.8
三菱	合計	3,250	738	3,988	7,319	0	70	70	121
	前年比	129.6	71.7	112.8	108.5		89.7	89.7	74.2
三菱ふそう	合計	0	0	0	0	2,033	780	2,813	5,163
	前年比					101.4	78.1	93.7	95.4
日産	合計	18,334	17,283	35,617	66,986	476	5,856	6,332	11,588
	前年比	90.9	93.3	92.0	92.8	91.2	112.9	111.0	125.1
スバル	合計	9,430	258	9,688	15,071	0	0	0	0
	前年比	90.6	135.1	91.4	75.4				
スズキ	合計	1,491	10,643	12,134	22,878	0	0	0	0
	前年比	101.1	113.2	111.6	111.4				
トヨタ	合計	57,307	59,336	116,643	213,795	1,914	12,778	14,692	26,567
	前年比	105.1	98.3	101.5	102.8	106.0	105.2	105.3	106.4
UDトラックス	合計	0	0	0	0	647	3	650	1,119
	前年比					97.6	150.0	97.7	102.0
	合計	21,355	1,324	22,679	39,967	65	0	65	109

(2)一次情報、二次情報

検索キーワードから見つけた Web ページには、お目当てのデータらしきものが掲載されていました。「これだぁ～！」と情報収集する前に、その情報は発信者（Web ページの作成者）が発信しているものか、それとも第三者が発信したものを転載しているものかを確認しておいた方が良いことがあります。



前者を「一次情報」、後者を「二次情報」と呼びますが、一般的に「二次情報」よりも「一次情報」のほうが信ぴょう性が高いといわれています。もし「二次情報」であった場合には、掲載されているデータの出典（引用元）へアクセスし、そのデータが確かなものかどうかの確認を行いましょう。「二次情報」は「一次情報」を探すためのヒントとして捉えておくことが良いでしょう。

出典が記されていない場合、その情報自体、またページの制作者側に問題があることが多いので、別のページを参照するようにしましょう。

(3) 1 次データ、2 次データ、3 次データ

データには、団体や組織、企業が直接収集した「1 次データ」、官公庁や関連団体が所有しているデータ等は収集を外部の専門機関に委託していることから「2 次データ」と、そして「1 次データ」「2 次データ」を加工し、閲覧者のニーズに合った形で編集されているデータを「3 次データ」と呼びます。

前述の POS システムで収集したデータは「1 次データ」となり、「2 次データ」と組み合わせて加工し、データ分析の目的を達成するために作成されたデータは「3 次データ」に該当します。

皆さんが企業で行うデータ分析にも、この考え方が必要になります。

4）データの入手方法

さて、ようやく探し出した Web ページから早速データを収集・・・、の前に少し考えておくことがあります。

データを入手する際の確認事項

1）利用制限の有無
閲覧者が自由に入手、利用してよいデータかどうかを
Webページ上で確認する

2）データ形式と、容量
・ファイルを開く前に形式を確認する
・容量を確認する

3）ファイルを開く前に、念のためウィルスチェックを行う

「そのデータ、勝手に使って良いの？」です。公開を主目的としている場合には問題になることがなくても、そのデータの所有者がデータの利用について「制限」を設けているケースもあります。その場合、Web ページ内に注釈がありますので、入手、作業を行なう前に必ず確認し、その範囲内で利用する、もしくは利用に際してデータの所有者に許諾を得るようにしましょう。

またデータの「形式」についても注意が必要です。データ分析を行う皆さんが探し出したページにはおそらく、膨大なデータが公開されているのではないでしょうか？その場合、Web ページに書き込まれているケースよりも、何らかのファイル形式で UP されていることが多いはずです。

Web ページによって異なりますが、丁寧なページではファイル形式やファイル容量を明示しているものもあれば、リンクをクリックしただけでダウンロードを行わせるページもあります。あまりに重いデータですと、インターネットの接続環境によってはダウンロードに時間がかかり、業務に支障がでることもあるので注意しましょう。

またファイルを開く前に、ダウンロードしたファイルはウィルスに感染していないか確認しましょう。Web ページの所有者が誰であれ、情報セキュリティの観点から必須の作業で

あることを認識しておきましょう。

5)データの引用

皆さんが見つけた Web ページの中には、お目当ての統計データだけでなく、そのデータの説明文や解説が掲載されています。学習したことを用いてレポート作成を行う際、「データ分析の報告書に使用したらすると見栄えも良いな」と、思わず Web ページの文章を「コピペ（コピーアンドペースト)」したくなるページもあるかと思います。Web 検索をおこなうと、必ずと言って良いほど上位に表示される Wikipedia では、調べたいテーマに関する詳細な説明が次から次へと見つかって、「おー、これこれ！参考にしちゃおう」とコピペしてしまうなんて気持ちになるかもしれません。でも、ちょっと待ってください。そこに大きな問題が潜んでいます。

ここではデータの引用における注意点について学びます。

皆さんは「引用」という言葉を聞いたことがあると思います。「コピペ」なんて言葉と混同している方も少なくないかもしれませんが、「引用」には正しいルールがあり、そのルールをきちんと守ることで「引用」と判断されるものです。逆に正しく引用していない場合、例えばコピペして終わりなんてことは著作権侵害となり、損害賠償請求の対象にもなりかねませんので、きちんと押さえておきましょう。

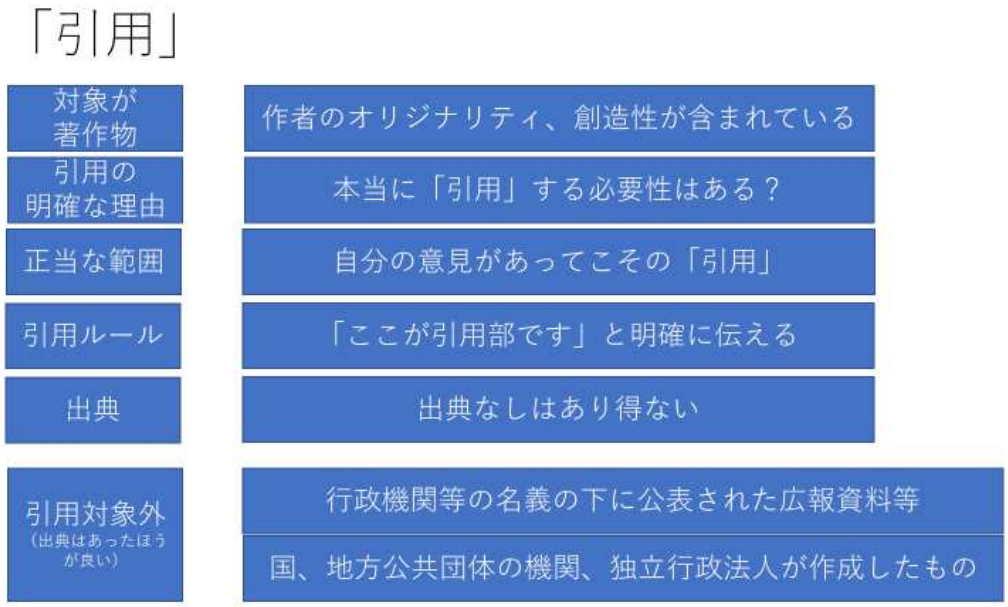
また「引用」する際にも「データの信ぴょう性」で述べたように、「引用」すべき Web ページなのか、そうではないのかが大変重要になってきます。前に学習した Web ページに掲載されている「データの信ぴょう性」とも関連するテーマとしてしっかり学んでおきましょう。

(1)「引用」とは

まずおさえておきたいのは、「引用」に関する正しい知識です。「引用」は正しく使用すれば薬に、間違った使い方をすると毒になります。そのために「引用」に関する正しい知識を掴んでおきましょう。

「引用」とは、“自分の考えを表現（研究、証明等）するために、「引用」の要件を満たしたうえで他人の著作物を副次的に使用すること”（出典：広辞苑）を指します。データ分析という行為においては、皆さんが行ったデータ分析を表現する（外部に対して発表等）ために、他人の著作物を副次的に使用行う行為になります。

この一文にはいくつかのキーワードが含まれていますので、そのキーワードについてまとめてみましょう。



まずは引用したいと思ったコンテンツ（文章、図、写真、動画など）が著作物であるかどうかを確認します。判断における最大のポイントは「著者のオリジナリティ、創造性が含まれているかどうか」になります。

次に「引用する必要があるかどうか」です。他人の著作物を引用することが自分の主張を補完するかどうかを判断しましょう。あえて引用する必要が無い場合、文章そのものが冗長になり、読む側にとって読みにくいものになりかねません。

さて次は、「適正な範囲（ボリューム）かどうか」です。一概に何行以下でないと NG と決めることは難しいですが、あなたの記述（表現）を超えるようなボリュームでは引用とは判断されません。あくまで引用であることが大切です。

引用することが決まれば、次に大切なことは引用する際の、「引用箇所の明示ルール」を守ることです。引用する際には、その引用箇所をはっきりと、あなたの著作物と区別できるように区別した記述を行うことです。

一般的には、文章であれば “”（ダブルクォーテーションマーク）や『』で該当箇所を括るといった方法があります。

最後はその引用がどの著作物から引用されたものかを表示する「出典の明記」です。出展の明記は（）等で、（出典：〇〇〇）と、出典元が明確にわかる必要があります。文章等の直後に入れることが一般的ですが、引用箇所が複数にわたる場合、「*1」といった記号を付記し、段落やページの最後にまとめて表記することでも可能です。

なお、データの中には引用には該当しないものもあります。例えば公的機関（国や官公

庁、独立行政法人等）が公表することを目的に公開しているデータ、また史実や事実報道等の文章などが該当します。

この場合、引用に該当しないから出典を記載しないという判断もできますが、データの信ぴょう性を担保していることを読み手に理解してもらうことを考えますと、出典をあえて明記するということも必要です。またあなたが後から読み返した際にも出典にあたる際の手間が省力化できるという利点もあります。

1.4.7. サイトアクセス時の危険性について

ページ検索を行う際に気をつけなければならない項目の一つに、「サイトアクセス時の危険性」があります。ページ検索では様々なキーワードを駆使して、数多くの Web サイトにアクセスを行うことになりますが、その分情報セキュリティを十分意識しておかなければなりません。ここでは「サイトアクセス時の危険性」について学んでいきましょう。

１）危険性の種類

一口に危険性と言ってもなかなかわかりづらいですね。具体的にどんな危険性が潜んでいるかイメージしてみましょう。

サイトアクセスの危険性

悪意を持ったページが存在することを意識する

(1)ウィルス感染
- ・クリックで感染
- ・ダウンロードしたファイルを開くことで感染

(2)インターネット詐欺
- ・偽サイト
- ・フィッシング詐欺
- ・ワンクリック詐欺　　他

大きく分けると、「ウィルス感染」「インターネット詐欺」に分けられます（※ウィルス感染からインターネット詐欺に陥れるサイトもあります)。

(1) 「ウィルス感染」にはページの中にウィルスに感染させるトリガー（引き金）が潜んでいて、『ここをクリック！』といったバナー（アイコン）や、画面に表示されるポップアップメッセージをクリックすることでウィルスが送り込まれるものです。気づかないう

ちにウィルスに感染していた場合、あなたのPCが接続されている情報ネットワーク経由で他のPCに感染させたり、あなたが入力した個人情報を収集したりと、できる限り早く対策をしなければなりません。

また Web ページにあるファイルをダウンロードし、開いた瞬間に感染するケースもあります。近年、「ランサムウェア」と呼ばれるウィルスが横行し、ダウンロードしたファイルから感染すると同時に、PCの利用者がシステムやデータにアクセスできないようにし、「利用したければ身代金をよこせ」といった脅迫メッセージを送りつけてくるものです。

これらの被害から身を守るためには、「クリックする前に」「ダウンロードする前に」、また「ファイルを開く前に」、もう一度安全を確認する行為が重要です。

「安全を確認する行為」としては、

・アンチウィルスソフトがアクティブになっているか
（パターンフアィルが最新のものにアップデートされていることが重要）
→危険なWebページを検知するシステムがONになっていると有効です。

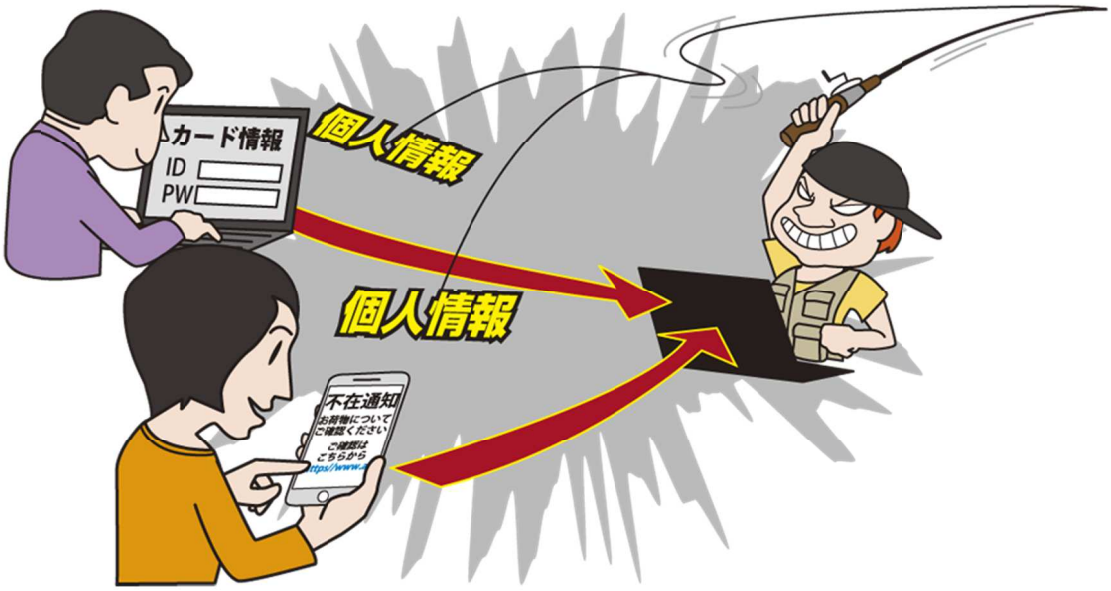
・ファイルをダウンロードする前に、ファイルの拡張子を確認する
→「.exe」といった実行ファイルは、安全が確認できなければダウンロードしない

・ファイルを開く前に、ウィルス検知をおこなう
→PCにインストールされているアンチウィルスソフトを使って、ファイルが感染していないか診断を行う

等があげられます。

(2) 「インターネット詐欺」は、Web サイト上で金銭や個人情報を不正に入手しようとする手口です。「偽サイト」は誰もが知っているような有名企業等のサイトと勘違いさせるようなページを作り、そのうえで金銭や個人情報を収集しようとするサイトです。とても巧妙に作られているケースが多いので、つい騙されてしまう人も後を絶ちません。

代表的な「フィッシング詐欺」ではスパムメール等から「偽サイト」にアクセスさせ、そこで個人情報をだまし取ろうとするものです。大手金融機関や e コマースサイトの Web ページに見せかけ、口座番号や暗証番号、また個人IDやパスワードを奪い、本人になりすまして現金を引き出したり、買い物をしたりする被害を受けてしまいます。



独立行政法人情報処理推進機構 発表 「情報セキュリティ 10 大脅威 2019」
第 2 位　フィッシングによる個人情報等の詐取　より参照

「ワンクリック詐欺」は、その名の通りワンクリックしただけで「契約完了」等の画面が表示され、身に覚えのない高額な金銭を請求されるものです。この詐欺にあった場合、提示された支払う必要のない金銭を支払ってしまう、また自分で何とかしようとしている間に誤って個人情報を渡してしまう（入力してしまう）などし、詐欺師に渡した個人情報を悪用されて被害に遭うこともあります。

被害に遭わないためには、「とにかく無視する」ことです。年々、手口が巧妙化し、振り込ませようとする（または個人情報を入力させようとする）ための様々な表現を使って脅してきます。（あなたがインターネット接続に使用している IP アドレスなどを表示する手口などが一般的です）
とにかく一旦、落ち着くことが大切です。

「電子消費者契約及び電子承諾通知に関する民法の特例に関する法律」の中で解説されている通り、消費者側が「購入（契約）する」ボタンをクリックし、再度消費者の購入（契約）意思を確認する確認ボタンをクリックして初めて、契約が成立するとされています。すなわち、ワンクリックで成立する契約は無効であるということです。これさえ知っておけば、仮に詐欺画面が表示されても必要以上に不安がる必要はありません。
ただし、その画面が表示された理由についてはあなた側にも問題があることも考えられますので、怪しい Web サイトや Web ページのアクセスはしないよう十分留意してください。

上記で紹介した以外にも、SEO ポイズニング他、Web は便利な反面、様々な危険も潜んでいます。ここで大切なことは必要以上に怖がることではなく、Web サイトアクセスの危険性を正しく認識し、対策を施し、正しい手順でデータを入手することです。

そのためにはインターネットユーザーとしての情報セキュリティに関する必要知識の習得と、そのアップデートを心がけましょう。

経済産業省の所管となる「独立行政法人情報処理推進機構（IPA）」の Web ページでは様々な情報セキュリティに関する情報を掲載し、個人、企業に対する啓蒙活動を行っています。基本的な情報リテラシから、緊急勧告的な情報までアップされていますので、ぜひ定期的に確認しておきましょう。

1.5.　メールによる送受信

1.5.1.　メールソフト

学生時代のデジタルコミュニケーションは、スマートフォン上で使い勝手の良い「LINE」、「Instagram」他の SNS アプリが主流となっているが、ビジネス・コミュニケーションにおいては、やはりメールソフトである。前節で解説した Web ブラウザを活用したデータ入手以外に、メールソフトを使用したデータの受け渡しの場面にも多く接することであろう。

ここでは、メールソフトの機能紹介、一般的な使用方法から、データのやり取りを行う際のファイル添付や大容量のデータ転送サービスについて紹介します。また、やり取りされるデータ（ファイル）の毀損についても説明します。

1）メールソフト

以上が、今年度開発分であり、全体構想（三ヶ年事業分）の 30%程度である。