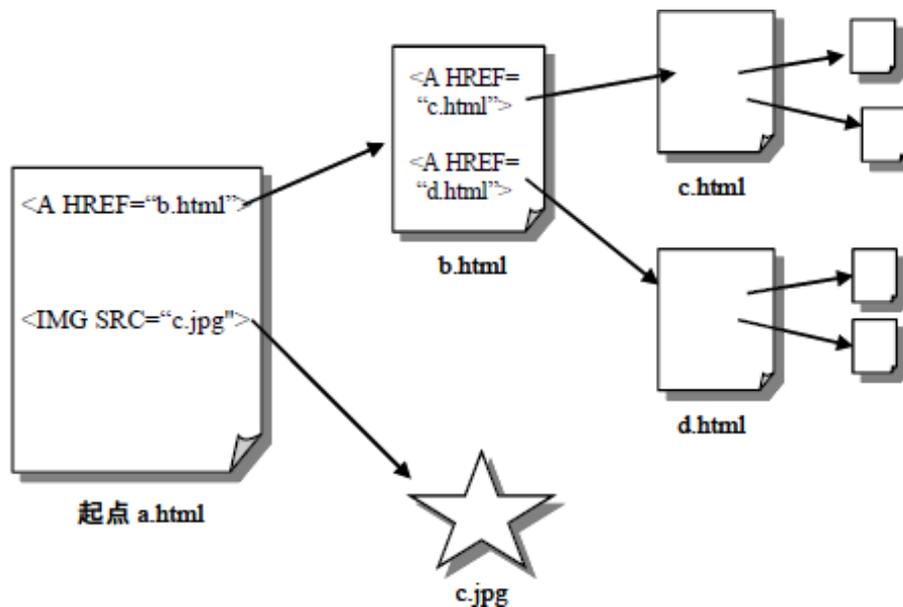


インターネット資料収集保存事業における収集の仕組みについて

1. 収集ロボットによる収集について

最初に、起点となるURL（サイト全体を収集する場合は、サイトのトップページ）にあるHTMLファイルを収集します。次に、その中のHREFタグやIMGタグを解析し、そのHTMLファイルからリンクされている文書、画像、音声、動画等の各種ファイルを収集します。以後、同様の動作を再帰的に繰り返し、最終的に収集対象全体のデータをダウンロードします（HTTPプロトコルにおけるGETコマンドによりダウンロードします）。起点となるURLが、リンクを含まない文書等のファイルであった場合には、起点の1ファイルを収集した時点でダウンロードは終了します。起点となるURLについては、当館で随時確認を行っておりますので、その都度ご連絡いただく必要はございません。



当事業における収集ロボットのUser-agentの名前は「ndi-japan」です。当事業の収集ロボットは、データ収集を制御するロボット排除規約に準拠していますので、収集ロボットを排除する設定がされている場合は、ロボット排除設定の解除をお願いいたします。なお、ロボット排除設定がされている場合、事前に連絡をさせていただいた上で、robots.txtファイル及びrobotsメタタグによるロボット排除設定にかかわらず、ご承諾いただいた範囲内のインターネット資料を収集する場合がございます。

なお、CGIプログラム等により動的に生成されるいわゆる「深層ウェブ」や、認証が必要なデータは収集できない可能性があります。ダウンロードと同時に再生されるファイル（ストリーミング配信）は技術的に収集できません。組織内のネットワークでのみアクセス可能なデータは、アクセス自体が不可能なため、収集できません。

2. 収集の負荷について

収集対象機関ウェブサーバの負荷軽減のため、ダウンロードの間隔を1秒以上あけます。収集開始日は、原則として当館が任意に定めています。収集日数は、収集対象のデータサイズやネットワークの状態に依存し、定まりません。収集頻度は、年数回を原則としています。

3. 収集範囲について

国立国会図書館による、貴機関のご許諾に基づく貴機関インターネット資料の収集にあたっては、収集から除外する範囲をご指定いただくことが可能です。収集の除外範囲は、同封の回答書別紙でご指定ください。

4. 収集に関する問い合わせ先

〒619-0287 京都府相楽郡精華町精華台8-1-3

国立国会図書館関西館電子図書館課ネットワーク情報第一係

電子メール：warp@ndl.go.jp