Web Archiving

Bearbeitet von
Julien Masanès

3 Selection for Web Archives

Julien Masanès

European Web Archive
julien@iwaw.net

3.1 Introduction

The selection phase (see Fig. 3.1) is a key phase in Web archiving. It takes place at the beginning of the entire cycle and has to be re-iterated on a regular basis. Preceding the capture phase for which it provides input and guidance, it comes just after the archiving and access phase of previous crawls if any, ideally taking into account issues and necessary changes the quality review phase has raised. It comprises three phases: preparation, discovery, and filtering that will be described in this chapter.
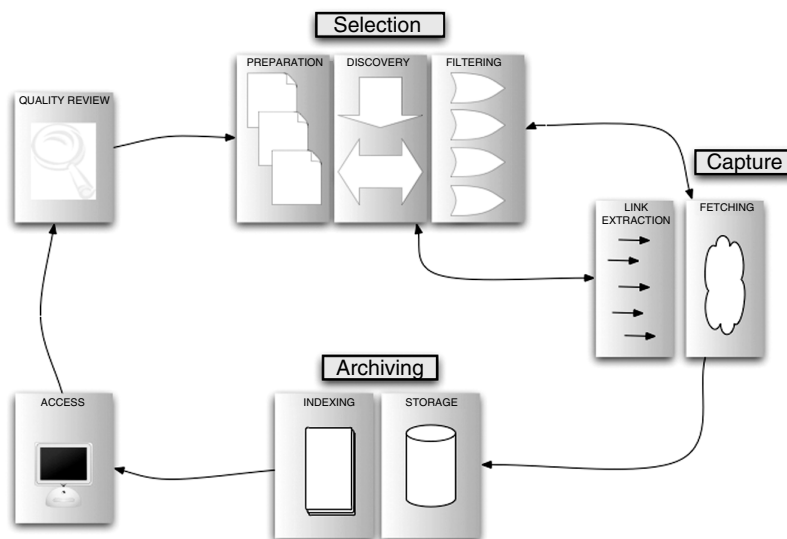
Fig. 3.1. The selection cycle, with its three phases (preparation, discovery, filtering), takes place before the capture, the archiving and quality review

The selection policy is the mark of each archiving institution. Choices made in this domain determine the type, extend and quality of the resulting institution's collection. But simply applying methods and practices developed for selection of printed material is not adequate. Web publishing is different enough from traditional publishing to require a wide revision of existing practices in this domain.

In this chapter, both the methodology and reflection on what selection means in the context of the Web will be presented with the ambition of contributing to such a revision. The chapter will cover the selection policy, the issues, and the implementation process of selection in the context of the Web.

3.2 Defining a Selection Policy

Building collections of Web material requires, when it becomes a regular activity, a general guiding document that defines the collection development policy. The benefits of defining such a policy for archiving institutions are the same as for printed material (Biblarz et al. 2001):

- It reduces personal bias by setting individual selection decisions in the context of the aims of collection building practice;
- It permits planning and identifies gaps in collection development and ensures continuity and consistency in selection and revision;
- It helps in determining priorities and clarifying the purpose and scope of each individual collection, and allows selection decisions to be evaluated by, for example, identifying what proportion of in-scope published material has been acquired;
- It can serve as a basis for wider cooperation and resource sharing.

Even if these benefits have been originally identified for collection of printed material (an electronic published resources by extension) the main principles remain for the web: avoiding personal bias or changes pertaining to a specific conjuncture hence providing continuity, defining priorities and allowing planning, positioning the collections in a larger archiving context to facilitate cooperation.

3.2.1 Target and Coverage

A collection development policy should describe at a high level and the goal driving the collection development. This comprises a description of

the context, the targeted audience, the type of access, and the expected use of the collection.

The collection's target, that is, the content to be archived, should be described in this context in general terms. This can be refined by defining inclusion's and exclusion's criteria. These criteria can be on quality, subject, genre, publishers like in traditional selection, but also on domains as defined by the Internet naming space itself. An importance difference to keep in mind for criteria adoption is their applicability on the web. It makes a huge difference in costs for instance if discovery as well as appraisal have to be made by human instead of automatically.

The concept of coverage or depth of collection has been widely used for books or serial to appraise collections, set ambitions and guide their developments. The five levels defined by the International Federation of Library Association (IFLA) are the following (Biblarz et al. 2001):

0 = out of scope
1 = minimal information level
2 = basic information level
3 = study or instructional support level
4 = research level
5 = comprehensive level

Collection can be appraised along several axis, the main one being subjects. Conspectus, by defining 24 divisions, 500 categories, and 4,000 subject descriptors can provide an outline of a collection that can be used as for systematic assessment of a library collection (Loken 1994).

The main problem when trying to apply this tool for web collections is the lack of reference to which comparing collection's completeness. This is partly due to the little number of existing institutions doing web content selection, compare to the numerous ones doing it for books or serials. For books or serials, one can easily find a large numbers of catalogs, bibliographic lists to refer to, not to mention the national bibliography made by national libraries, which provides an almost complete survey of the printed material for most countries.

This is also due to the qualities of the web as a publishing medium, which makes this type of rigid framework usually hard to apply. Traditional publishing professionalism and structuring for a well-established market prepared the ground for librarians' effort to organize printed material. As it will be discussed in section "Limitations" (implementing a selection policy) the nature of the Web makes implementation of a collection policy quite different from a traditional one.

Two main differences should be emphasized here: the connected nature of the information space and the lesser role played by professional publishers with the multiplication of content producers that goes with it.

Link connectivity deeply structures organization of information on the Web. Selection morphs from a human selection of discrete and stable units (books or serial) to a more versatile and dynamic selection of paths to be followed with certain depth and time. The topology of the Web plays a key role in both the discovery and the capture of content on the Web. It indeed tends to replace the well-structured organization of selection of the book era where publishers, collections, disciplines were natural lines along which selection was organized.

There are of course fewer differences for human-driven selection than for automated ones. However, the multiplication of content publishers, the variety of publication's forms and frequency, the nature of discovery and authority on the Web requires adapting traditional practice significantly.

For an example of a selection policy closest as possible to the traditional model, see the NLA's selection policy (National Library of Australia, 2005).

3.2.2 Limitations

Whereas traditional acquisition policy had mainly to deal with financial limitations (for acquisition, processing, or storage) web archiving is also directly and permanently hindered by technical difficulties for capturing content. Different types of technology challenge current capture techniques: the hidden web (see Chap. 5 on hidden Web archiving by Masanès 2006a), streaming content, highly interactive content etc. There are hence hard limits to what can actually be archived. There is also an inherent limitation to server-side archiving (the main archiving methods) that can only capture functionality that are supported by client-side code. The development of AJAX web programming style based on content exchange between the page and the server without reloading the page can augment the amount of material not captured by crawlers. These limits have to be included in a selection policy when possible, as they will impact the resulting archive quality. Here is for example the list of exclusion of NLA's policy. Although we could not say whether this had been the main or only one of the reasons behind some exclusions (cams, datasets, games), technology would have been a challenge in these cases anyway.

– Cams (websites employing a Web camera that uploads digital images for broadcast);
– Datasets;
– Discussion lists, chat rooms, bulletin boards, and news groups;
– Drafts and works in progress, even if they otherwise meet the selection guidelines;
– Games;
– Individual articles and papers;
– News sites;
– Online daily newspapers for which print versions exist;
– Organizational records;
– Portals and other sites that serve the sole purpose of organising Internet information;
– Promotional sites and advertising;
– Sites that are compilations of information from other sources and are not original in content;
– Theses (the responsibility of universities and the Australian Digital Theses Project).

3.2.3 Gathering Patterns

Building Web collection can either be done on a continuous basis or through campaign or snapshots.

Examples of these campaigns are elections sites acquisitions or snapshot domain crawls. Although archiving campaigns can change on emphasis or thematic, they should be done accordingly to the collection development policy. Conversely, the collection development policy should describe, when possible, the campaign patterns to make sure the end result is consistent with the overall aim of the collection development. A campaign pattern description should at minimum comprise the trigger(s) (calendar, events, etc.), duration(s) of the campaign as well as the possible bridges to be established between campaigns.

Here is a very simple example of such campaign pattern description:

Start national domain snapshots every three months, with a campaign duration of 60 days and by using as entry points, list all the domains found in the previous campaign.

3.3 Issues and Concepts

3.3.1 Manual vs. Automatic Selection

A recurrent theme in the literature on Web archiving is somehow simplistic opposition between manual selection and bulk automatic harvesting allegedly considered as unselective. The former is misleadingly supposed to be purely manual whereas the latter is similarly falsely considered as comprehensive. We prefer to insist on the fact that Web archiving always implies some form of selectivity, even when it is done at large scale and using automatic tools.

There are two levels at which this selectivity and the determinism of automatic tools takes place: discovery and capture of material. Comprehensiveness as opposed to selectivity is a myth as Web's size and versatility make it impossible to discover and to capture every possible instantiations of content for all possible readers. Actually, there is a default selectivity of large-scale crawlers in term of the extent, depth, and time at which they crawl sites, all these in turn being dependent on resources used, capacity to extract links, queuing method, politeness to servers, entry points used, etc. We prefer to use in this book the term holistic archiving, defined as archiving made by open crawls using link extraction for discovery.

On the other end, manual selection of Web documents rarely happens without requiring utilization of automatic discovery tools like search engines. See a detail modeling of user/machine interaction for IR in general and Web search in particular in Ellis et al. (1998). It should also be noted that these tools add an access bias (ranking methods) to the crawl bias (see for instance Introna and Nissenbaum (2000) and Vaughan and Thelwall (2004).

And even in the case where discovery would entirely be done manually, capture is most of the time done with tools based on link extraction (site copiers). Here again, one has to be aware of the fact that these tools always have at least embedded capture bias like definition of scope, implicit or explicit exclusions of content by format type, prioritization of capture, resources constraints (hardware, bandwidth), etc. This underlying determinism of web capture has a sufficient impact on the final resulting collection not to be underestimated (see Chap. 4 Roche 2006).

3.3.2 Location, Time and Content

As we all know, and despite the fact that this is totally counter-intuitive, there is no such a thing as reference to objects on the Web. URLs provide references to locations, not objects and applying a selection policy in a space only structured in terms of location is more challenging than one could think at first glance. To take a familiar example, it is like walking in an open stacks library and selecting books only by their location (shelves) while objects can be moved or removed from time to time. But going straight to the shelf containing medieval history books, if this is what one is passionate about, is only one of the two main possible means for a reader of selecting book in a library. Most of the time, she/he would use the other one, the catalog. And catalogs handle objects identifiers, for which they provide location.

Web references on the contrary, handle locations first, then objects. The difference does not only come from the order, but also from nature of this relation. In one case (the library), the relation between the object and its location is maintained by the library itself which guarantees it will work whatever happens to the original publisher (and we know that libraries in general last longer than publishers). Whereas on the Web the relation between the object and its content depends on the publisher. Moreover, it actually also depends on the publisher's technical ability and permanent use of resources needed to serve content online. This is what can be called permanent publishing.

It would be misunderstanding the real nature of the Web to think of this characteristic as a shortcoming. It indeed offers the ability to create a space where each defined location (like a domain name) is a source of possibility instead of being a placeholder for fixed content, where the producer will have the possibility to propose, update, change, and remove content, which is the very nature of publication after all. A location is then a way to connect to someone's or something's stream and this can become a criterion for selection.

A consequence of this is the introduction of a new temporal dimension in the archiving process. Actually, archivists are more used to deal with this than librarians are. Their activity has always been closely linked to the document lifecycle whereas librarians have mostly been working on stabilized (published) content. For the Web, as contents update or removal can occur at any time with great facility, this temporal dimension has to become a core component of any process of archiving. The fixation of a particular state of content, which used to be done only by the publisher, is now also made by the archive.

This is a new responsibility for preservationists in general, and it alters significantly the usual relation between location time and content in traditional archiving as well as actor's roles in this process. By capturing and storing content on its servers, the Web archives removes completely the resources dependency to its original creator and eliminates the virtuality of any Web location, "freeze" the resource in time and therefore reduces its dual nature to its content aspect only. This, which used to be done by publishers and printers for printed material, falls under the responsibility of the archive for Web material.

3.3.3 Hubs and Targets

Even if little has been done on selection with archiving as a goal, a significant precedent effort was made for creating reference list or subject gateways pointing to selected resources on the Web. It is interesting to note that the dual nature of Web references is sometime pulled on one side or the other. The location aspect is certainly more important for resources that contain themselves mostly reference and links to other resources (hubs, also called webographies, etc.). This is also the case where the resources are valued for providing up-to-date information. They are valued and selected as locations where updated and reliable information on a subject can be found, that is, locations where change can and has to happen.

Some resources on the contrary are valued for their content itself. This is usually the case for smaller piece of content, pages, single documents, and/or content dependent on a specific time (events).

The graph theory offers concepts that can prove useful to characterize the two types. Considering the graph formed by pages (nodes) and links (directed edges) it is easy to calculate the in-degree of a node (number of nodes that link to it) or its out-degree (number of nodes it links to). From a pure structural point of view, we can expect a hub to have a high out-degree, and a low ratio content/out-link (content can be measured by the quantity of information the node contains for instance). On the contrary a target can be expected to have a low out-degree and overall an important ratio content/out-link. If this target is important, it can be expected to have, in addition to this, a high in-degree.

An iterative definition of hubs and authority can be found in Kleinberg (1999) and can be summarized as:

Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs.

Both concepts can be applied for Web archiving with slight modification. We will use the term "target" instead of authority in this book, in order to remain neutral with regard to the "authority" of each individual node selected. An archiving policy can target content for a wider range of reason than just authority. Important to note is the nonexclusivity of the two types: a hub can also be a target for a selection policy as it can provide content as well as referral information on other resources. Hubs and target can be selected and linked to by subject gateways. Targets are vulnerable to time as time entails, at least potentially, change, whereas hubs are pointed to by subject gateways because of the same reason, as they offer a good chance to be changed if needed (updated).

One could conclude rapidly that hubs are of little interest for Web archiving selection policy, which should concentrate its effort on the first type, vulnerable to time. But this is not the case, at least for two reasons. The first one is that hubs can attest of relations like valuation, endorsement, etc. that could prove to be of great interest and deserve archiving for themselves. The second reason is that even when they are not targets, they certainly are a mean for finding targets. Hubs are indeed a tool of choice to implement a web archiving policy.

3.3.4 Entry Point and Scope

The term hub and target have been defined from a pure structural and content-quality perspective. From a practical point of view, we need to introduce two other terms related to how a selection policy can eventually be implemented. The first one, entry point (EP) also called seed, could be confused with the concept of hub. An EP is defined as the first node from where path to other documents will be found in a crawling process. As they both have to do with out-linking, hubs, and entry point often tend to be confused and indeed, most EP are usually hubs themselves. But this is not always the case. An example of this is the site's homepage, often considered as the EP for a site crawl whether it is a good hub beyond the site itself or not. The related concept of scope can be defined as the extent of the desired collection, delimited by a set of criteria. Criteria can be topological (the Italian domain), thematic (sites related to biology), based on genre (blogs) time (site stale since the last 2 years), etc. Each time a new path is discovered from the EP or pages linked from it, it has to be evaluated to see if it fits in the scope or not. To be operational, scopes have to be defined in a way that enables direct and possibly automatic verification. If not, a systematic human evaluation is needed for each new link discovered. If the selection policy is applied at the site level, this will only be necessary when a link to an external site is discovered.

3.3.5 Level of Application

Traditionally, selection policies have implicit application levels, defined by the physical shape of their object (obviously, a selection policy applied at the level of book pages would have been nonsense). Stating the accepted types of documents (books, serial but not reports for instance) was sufficient in this environment. With the Web, this is no longer the case. There is no obvious level of selection and sometimes, levels are difficult to delineate (see for instance Thelwall 2002; Halavais 2003). It has been argued that the appropriate level for analysis in political science is the Web sphere, defined not simply as collection of websites but as

A set of dynamically defined digital resources spanning multiple websites deemed relevant or related to a central event, concept or theme, and often connected by hyperlinks. The boundaries of a Web sphere are delimited by a shared topical orientation and a temporal framework" (Schneider and Foot 2005).
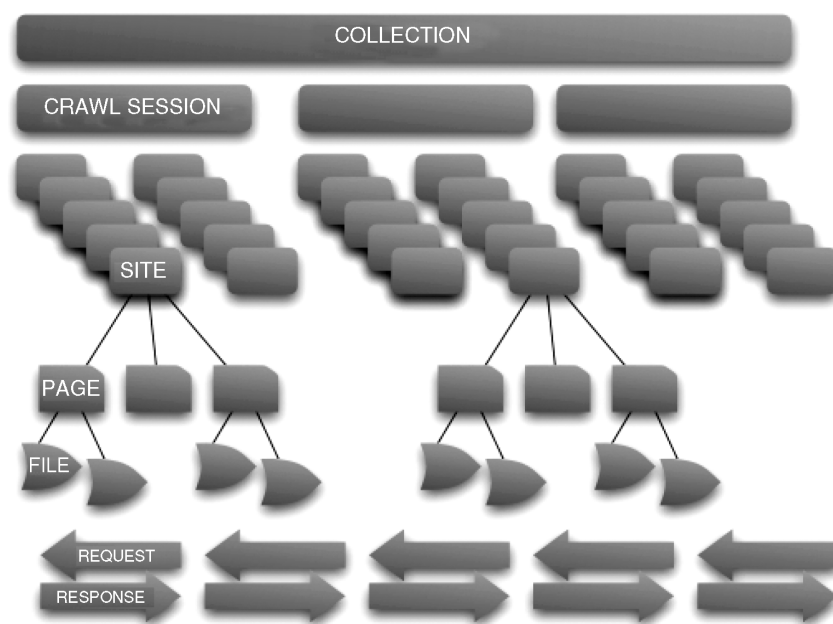
Fig. 3.2. Levels of information in Web archiving, from the request/response exchange to the site level for Web levels, from the crawl session to the collection for the archive's level

For selection, at least two working levels have to be considered: the page and site level. The page level corresponds to the immediate experience of web users and can therefore always be specified by human as well as by tools (browsers). It includes the skeleton (usually an html page) together with its embedded elements when rendered by the browser that is, images, style sheets, script files, etc. It can also consist of a non-web document and be rendered using helper applications (like a PDF document for instance).

Although it is used in many research studies (McMillan 1999), the site level is more difficult to define. The intuitive notion of a site refers to a related set of resources, sharing a same creating entity. The network notion refers to the host or web server that is, the machine serving the content of the site. Finally a purely topological notion of a site can be defined as a naming space section (a domain name for instance).

Confusion arises from the fact that these three levels are not clearly delineated internally. An entity can be an organization, a department, a person or even a project with its own site. A single machine can host several websites (and conversely, a site can be hosted on several machines). There is neither a strong naming convention for sites. A site can be located at the level domain (netgramme.org) at the sub-domain level (zone.netgramme.org) or even at the directory level (netgramme.org/blog). The situation worsens as the three levels get confused one with the other. This flexibility in possibilities (and in actual usage) is typical of the web.

As selection policies are usually content driven, they should focus on the first and third level (discarding the machine level). Defining target sites can be done by assessing in each case the appropriate entity level. If the collection targets neuroscience related material, neurobiology research laboratory's sites are certainly more interesting than entire university Websites for instance. The hierarchical nature of naming conventions can often help here as long as they have been applied for the site construction. Therefore when identified, the most characteristic path can be passed over to crawler that can deal with it by getting only the content further down in the hierarchy. For instance, if the site has been identified under a specific directory (netgramme.org/blog) than the crawler can limit the crawl to content that is under this directory (however, deep). Note that the naming hierarchy goes from right to left for the domain name and left to right for the path (see Fig. 3.2).

3.4 Selection Process
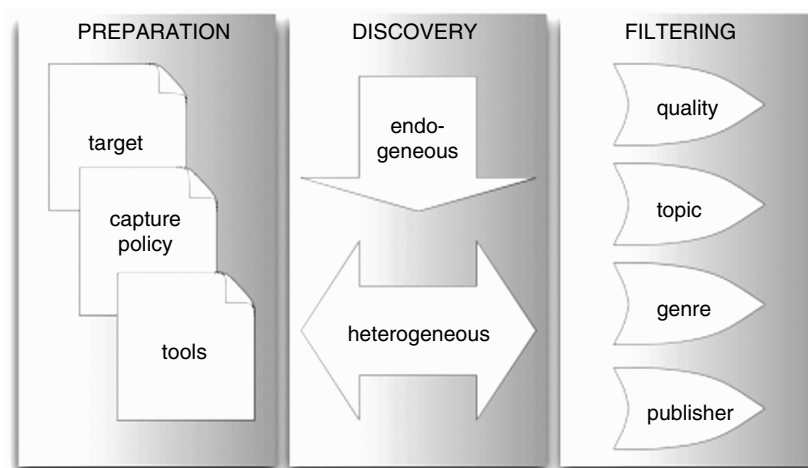
Fig. 3.3. The phases of the selection process: (1) preparation with its main output (the target definition, the capture policy and the list of tools to be used), (2) endogeneous and heterogeneous discovery and, (3) filtering according to quality, topic, genre or publisher

The selection process can be divided in three main phase: preparation, discovery, and filtering (see Fig. 3.3). Although these phases can occur in a sequential order or can be mingled together to some extent, we will present them as logically distinct for sake of explanation.

3.4.1 Preparation

This phase is a key for the success of the whole process and should not be underestimated in terms of time as well as resources required to perform it successfully. The main objective of this phase is to define the collection target, the capture policy and the tools for implementing it.

   For topic-centric as well as domain-centric collections, input here is mainly required from domain experts (references, librarians, archivists, scholars) that have to define what the target information space is, how it can be characterized in extension and granularity, and which frequency of capture will be applied.

   The definition has to be precise enough to be implemented. Whether the discovery and filtering phases will be implemented manually or automatically makes a huge difference with regards to what "precise" means here.

Here are two examples, one with a strong human input in the discovery phase, the other where crawling will entirely perform discovery.

Example 1: Presidential election campaign collection:

− The archiving campaign will start 3 months before and end 1 month after the election date;
− All party, campaign, blogs sites of each official candidate will be archived entirely each week during the capture campaign;
− Main analysis, commentary and humorous website entirely dedicated to the elections will be archived every month during the capture campaign;
− Individual articles from the national and regional newspaper's websites will be archived once;
− The presidency website will be archived each month during the archiving campaign.

Example 2: National domain capture:

− Capture all French public sites on a bi-monthly basis, French sites being defined as sites from the .fr TLD or sites in generic TLD that are hosted by servers in France (based on the telephone number provided for DNS registration);
− A seed list of 12 general directories is provided to initiate the first crawl. Next crawls will start from the list of sites discovered during the previous crawl with complementary list of missing sites manually selected.

It is quite obvious that appraising what a campaign site of a candidate is, as in the first example, requires human judgment. Finding as well as filtering sites along this criterion will require manual processing. Whereas, in the second example the discovery and filtering criteria can be directly interpreted by robots. The preparation phase also requires defining which tools will be used during the discovery process. Four categories of tools can be used:

3.4.1.1 Hubs

Hubs can be global or topical directories, sites or even single pages with important links relevant to a given subject. These hubs are maintained by human, and often provide a valuable source for identification. Their reliability, freshness as well as their coverage has to be assessed on a periodic basis. When possible, direct contact with the person(s) in charge of a hub can be fruitful to better understand how their input can be used. Monitoring these hubs during the capture campaign as is necessary to ensure, they

remain relevant and exploits their input. This can be facilitated if they provide RSS or Atom threads.

3.4.1.2 Search Engines

Search engines can facilitate discovery of relevant material as long as precise enough query terms can be defined. Utilization of specialized search engines, when possible, can greatly improve relevance as well as, sometime, freshness of results. When the topic is closely related to a specific event, one should expect search engines to find relevant information only with a certain delay, which limits their usefulness in this type of capture. It can be helpful to define a list of queries as well as a list of search engines to use during the preparation phase. A periodicity of query and/or a mechanism to get updates (query-based RSS feeds or agent that filter new results) is also worth defining.

3.4.1.3 Crawlers

They can be used to extract links from already known material in a systematic manner. This can be used for exploring proximal environment of a given set of EP.

3.4.1.4 External Sources

Non-Web sources can be anything from printed material to mailing lists etc. that can be monitored by the selection team. They should be used when possible as they often provide fresh resources as well as different directions for the collection. Here again a monitoring process as to be put in place as this can easily become time-consuming and yield too little compared to time invested. It should be noted that, depending on the external sources authority, an item's citation by this source could, by itself, become a reason to select it.

At the end of the preparation phase, the following output should be available:

– The collection's target description;
– The capture policy, including the level of application, the frequency and extension of the capture;
– The list of tools that will be used for discovery and capture with a description of how they will be used.

3.4.2 Discovery

The main goal of this phase is to determine the list of entry points that will be used for the capture as well as the frequency and scope of this capture. It should be noted that there is a quite clear cut between discovery and the crawl itself for collection done manually, even if the list of entry points can be updated based on links discovered during the crawl. For automatically built collections, this difference is blurred by the fact that most of the discovery occurs during the crawl itself by links extraction. However, we can differentiate for both methods, "endogenous" discovery made from the exploration of EP's and crawled page's linking environment, from "exogenous" discovery that will result from the exploitation of hubs, search engines, and non-Web sources. Where "manual" collections mainly rely on exogenous discovery, "automatic" collections rely mostly on endogenous discovery for building collection.

Endogenous discovery takes advantage of the link structure of the Web to traverse and find new material. There is evidently a good chance that sites or more generally resources linked together deserve to belong to the same collection, as links are usually the expression of a semantic or topical relation (on the centrality of links see for instance Jackson, 1997). We will see in the next section (Filtering) how to qualify this topical proximity using textual content or linking structure. Let's just note here that related content can sometime be connected not directly but through several hops. This is for instance the case in competitive environment. Competitors will hardly directly link each other. Traditional citation analysis has studied this phenomenon extensively and showed that utilization of co-citations (two papers with no direct reference to each other will be both linked from or link to a common third paper) is a way of overcoming this problem (see for instance Garfield 1979).

The same can occur also at a macrolevel, the community level, where communities are strongly interconnected (which permits good discovery within the community), but loosely if at all connected across communities. Thus, the community forms a closed sub-graph where an endogenous discovery process can be trapped. In this case, it is either necessary to permit several hops with no filtering rules applied to "tunnel" out of these sub-graph (Bergmark et al. 2002), or use heterogeneous strategies (like insertion of meta-search results (Qin et al. 2004) in the discovery process). Finally, let us note that from a discovery point of view, linked resources belonging to different websites bear more value than those within the same site. They indeed bridge different information space, possibly belonging to different publishing organizations. Eiron and McCurley show that 33% of links are of this type (Eiron and McCurley 2003).

Heterogeneous discovery does not share this problem of sub-graph trap, as it uses sources that are not (or supposed not) to be linked to any specific community or portion of the hypertext graph. It, however, entirely depends on the type, quality and usability of the sources used: hubs, search engine or non-Web sources. The usefulness of the first type (hubs) obviously depends mainly on the quality and freshness of the source. Using the second one (search engine) permits to exploit the large and neutral Web exploration artifacts that giant search engine crawls represent. The difficulty is then to be able to query efficiently their huge inverse index. However, it has been shown (Lawrence and Giles 1998) that search engine coverage is relatively small and that there is little overlap between them. Using several of them (directly or through meta-search engines) is required to achieve a better coverage.

Non-web sources require specific monitoring, adapted to each case. The paper press, for instance, can be a rich source for an event-oriented collection either directly when websites are mentioned in articles, or indirectly when names or specific words can be found and use for a search.

Heterogeneous and exogenous discovery are not completely separated and a blend of the two can result in better results (Qin et al. 2004).

When entry points are discovered, a frequency and a scope of capture have to be assigned to them. This can be done individually or based on grouping of EPs. It is usually either done at the collection or capture campaign level, by defining one or several profiles of captures.

The usual frequencies are "once only", weekly, monthly or every x months. It is rare that capture have to be done on a daily basis or even several times a day. This can however be necessary for online news sites (see for instance Christensen-Dalsgaard 2004) or sometime for event-related captures like the September 11 Web Archive (Schneider et al. 2003).

The scope of capture is also important to define. As mentioned earlier ("Level of application") defining boundaries in the web is not simple. However, the page and the site level (defined as the domain, sub-domain or directory location) can be used, as they are directly understandable by crawlers. The units can be either in the entry point list or discovered from them. They can be used for defining the boundaries of capture in a restricted or extended manner. The restricted scope is to limit the capture to a specific page or sites that

Following any links from the entry points could result in an endless crawl of the entire Web. It is therefore necessary to shape this discovery process accordingly to the selection policy.

3.4.3 Filtering

The filtering phase's main goal is to reduce the space opened by the discovery phase to the limits defined by the selection policy. As already mentioned, if this phase can be logically distinguished from the others and particularly the previous phase of discovery, they can, in practice be combined one with the other.

Filtering can be done either manually or automatically. Manual filtering is necessary when criteria used for the selection cannot be directly interpreted by automatic tools or robots. This can be the case when high level characterization, subjective evaluation, and/or external knowledge is needed. As costs associated with individual selection and updating of list of resources by humans are high, there is a strong incentive to find ways of replacing or enhancing the efficiency of this scarce resource. Furthermore, the frontier between what is and is not interpretable by robots is highly dependent on technological evolution, and significant progress can be expected in this domain. It is, after all, a question of exploiting humanly generated intelligent information like words and links in pages in an automatic manner. Strong correlation between human appraisal and what can be assessed from structural properties reflecting this collective intelligence of the web have been established for instance by Masanès (2002) in the context of web archiving.

But in some cases, human input is still needed. Consider for instance our first example of collection policy, defining what the "main analysis, commentary and humorous website" requires both high-level characterization ("analysis", "commentary", "humorous") and subjective evaluation ("main"). "All party, campaign, blogs sites of each official candidate" also requires knowing which are the official candidates, that is, an external knowledge about the campaign itself. When direct human input cannot be replaced, it can be greatly optimized by using an appropriate and ergonomic presentation of items to be filtered. This includes for instance contextual information, visualization tools, and maps (see Cobb et al. 2005, for instance).

It is also important to define the appropriate level at which manual filtering has to occur to avoid duplicate evaluation (the higher the better to save time).

Several evaluation axes can be used, alone or in combination, for manual selection:

3.4.3.1 Quality

This comprises an appraisal of authority and credibility for secondary resources (like, in our example analysis and commentary websites of the campaign) and relevance and authenticity for primary resource (sites of political party).

3.4.3.2 Subject

A subject can be delimited along the traditional scholarship disciplines (biology, geology, etc.), or according to a specific event, person or organization, or any object in general, which will be envisaged from various points of view (like in our example, the elections). Here again, primary as well as secondary websites are to be considered for inclusion in the selection policy.

3.4.3.3 Genre

Web genre is institutional website, blogs, personal pages, forums etc. This can either be the main selection criteria for genre studies or an additional criterion (as blogs in our example). Genre have been studied in the context of the Web to see how they replicate or diverge from the genre in the printing world (see Crowston and Williams 1997), or how they can be automatically identified (Rehm 2002) and several genres have been studied like newspaper (Eriksen and Ihlström 2000), in homepages for instance Ryan et al. (2003) or FAQ (Crowston and Williams 1999).

3.4.3.4 Publisher

Traditionally publishers' reputation or specializations have been used to guide selection of printed material. It is often difficult to determine the publisher of a website and only very regulated top-level domains TLDs offer homogeneity of publisher's type, like the .mil for US military sites. Using the publisher or site owner as a basis for selection hence requires most of time detailed analysis of the site as no guarantee exists that claims of identity are legitimate on the Internet. The DNS information can be used to find out who is renting the domain as it requires registration of name and contact information of the technical and administrative contact. But depending on the TLD and the way it is managed, this can be either well quite complete or very limited.

3.5 Documentation

Whatever criteria are used for manual or automatic selection, it is necessary to document carefully the selection process. As we have seen previously (see Chap. 1 of this book Masanès, 2006b), web archiving can only achieve sampling of instantiation of content. As time goes, the original context of the sample is lost and no clue will remain for researchers to understand what the archive represents. To limit this, it is absolutely necessary to document each aspects of the selection process in order to provide elements of assessment for the future. This has to be done for the various phases outlined in this chapter (preparation, discovery, and filtering).

For the preparation phase, the main aspects to document are:

− The target;
− The capture policy and infrastructure (this comprises the technical capacity, software used, priority, politeness etc.);
− The tools used (name, regularity and context of use, staff, etc.).

Documenting the discovery and filtering is even more important as this will "tell" why a piece of content is or is not is in the collection. When possible, this should be documented at the item level. Keeping a list of URI that were discovered but filtered out can for instance be useful to later understand how the collection was built and therefore, what it represents compare the live web. How endogenous discovery was used is also important to document to be able to reconstruct path that were followed and map those that were not.

3.6 Conclusion

Selection is a key issue for web archiving. Manual selection can prove useful for a specific community and/or goal, where high-level assessment of items is necessary. As long as they cannot be made by robots, human selection has to be used, and it is necessary to organize it optimally. But even for holistic crawls, there is a level of selectivity and prioritization that has to be acknowledged and organized. We have presented in this chapter an analytical view of this process selection that tries to show that the two approaches share the main elements of this process, even if their relative importance and their use are different.

References

Bergmark, D., Lagoze, C., & Sbityakov, A. (2002). *Focused crawls, tunneling, and digital libraries*. Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries, Roma, Italy

Biblarz, D., Tarin, M.-J., Vickery, J., & Bakker, T. (2001). Guidelines for a collection development policy using the conspectus model. *International Federation of Library Associations and Institutions, Section on Acquisition and Collection Development*

Christensen-Dalsgaard, B. (2004). Web Archive Activities in Denmark. *RlG DigiNews*, *8*(3)

Cobb, J., Pearce-Moses, R., & Surface, T. (2005). *ECHO DEPository Project*. Paper presented at the 2nd IS&T Archiving Conference, Washington, USA

Crowston, K. & Williams, M. (1999). *The effects on linking on genres of Web documents*. Paper presented at the 32nd Hawaii International Conference on System Sciences (HICSS-32), Hawaii, USA

Crowston, K. & Williams, M. (1997). *Reproduced and emergent genres of communication on the World Wide Web*. Paper presented at the 30th Annual Hawaii International Conference on System Sciences (HICSS-30), Wailea, USA

Eiron, N. & McCurley, K. S. (2003). *Locality, hierarchy, and bidirectionality on the Web*. Paper presented at the Workshop on Web Algorithms and Models

Ellis, D., Ford, N. J., & Furner, J. (1998). In search of the unknown user: Indexing, hypertext and the World Wide Web. *Journal of Documentation*, *54*, 28–47

Eriksen, L. B. & Ihlström, C. (2000). *Evolution of the Web news genre – The slow move beyond the print metaphor*. Paper presented at the 33rd Hawaii International Conference on System Sciences (HICSS-33), Hawaii, USA

Garfield, E. (1979). Mapping the structure of science. *Citation indexing: Its theory and application in science, technology, and humanities.* NY: Wiley.

Halavais, A. (2003). *Networks and flows of content on the World Wide Web networks and flows of content on the World Wide Web*. Paper presented at the International Communication Association, San Diego, USA

Introna, L. D. & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *Information Society*

Jackson, M. H. (1997). Assessing the structure of communication on the World Wide Web. *Journal of Computer-Mediated Communication*, *3*(1), http://www.ascusc.org/jcmc/

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computer Machinery*, *46*, 604–632

Lawrence, S. & Giles, C. L. (1998). Searching the World Wide Web. *Science*, *280*, 98–100

Loken, S. (1994). The WLN Conspectus. *Cooperative collection management: The conspectus approach* (p. 107). New York: Neal-Schuman

Masanès, J. (2002). Towards continuous Web archiving: First results and an agenda for the future. *D-Lib Magazine*, *8*(12)

Masanès, J. (2006a). Collecting the hidden web. In J. Masanès (Ed.), *Web archiving.* Berlin Heidelberg New York: Springer

Masanès, J. (2006b). Web archiving: issues and methods. In J. Masanès (Ed.), *Web archiving.* , Berlin Heidelberg New York: Springer.

McMillan, S. J. (1999). *The microscope and the moving target: The challenge of applying a stable research technique to a dynamic communication environment*. Paper presented at the 49th Annual Conference of the International Communication Association (ICA-99), San Francisco, USA

National Library of Australia. (2005). Online Australian publications: Selection guidelines for archiving and preservation by the national library of Australia

Qin, J., Zhou, Y., & Chau, M. (2004). *Building domain-specific web collections for scientific digital libraries: A meta-search enhanced focused crawling method*. Tuscon, AZ, USA

Rehm, G. (2002). *Towards automatic Web genre identification a corpus-based approach in the domain of academia by example of the academic's personal homepage*. Paper presented at the HICSS-35

Roche, X. (2006). Copying websites. In J. Masanès (Ed.), *Web archiving.* Berlin Heidelberg New York: Springer

Ryan, T., Field, R. H. G., & Olfman, L. (2003). The evolution of US state government home pages from 1997 to 2002. *International Journal of Human.-Computer Studies*, *59*(4), 403–430.

Schneider, S. M., Foot, K., Kimpton, M., & Jones, G. (2003). *Building thematic Web collections: Challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive*. Paper presented at the 3rd Workshop on Web Archives (IWAW'03), Trondheim, Norway

Schneider, S. M. & Foot, K. A. (2005). Web sphere analysis: An approach to studying Online action. In C. Hine (Ed.), *Virtual methods: Issues in social science research on the Internet.* Oxford, UK: Berg.

Thelwall, M. (2002). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university web sites. *Journal of the American Society for Information Science and Technology*, *53*(12), 995–1005

Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing and Management*, *40*(4), 693–707