

Webinator and Search Appliance

Basic Concepts

- Web-based admin
 - **dowalk**: Administration
 - **search**: User search
 - **webmin**: Search Appliance system admin
- Profiles
 - Walk settings
 - Search settings

Installation (Webinator)

- Uninstall first
- **webinator.exe license.upd**
- Upgrade notes
- License agreement
- Install dir
- CGI bin dir
- Doc root dir
- Password

Installation (Webinator)

- Install Taxis ISAPI
 - IIS 5 or 6
 - ISAPI dir
 - Port
- Install
- Finish
 - Restart IIS
 - Open the admin interface

Uninstall (Webinator)

- Control Panel
 - Add or Remove Programs
- Restart IIS now
- Finished
- Ensure all files deleted
 - Profile dataspace

Crawl Settings

- What to crawl
- How to crawl
- When to crawl
- Data manipulation
- Resource limits
- Authorization
- Miscellaneous

What to crawl

- Starting points
- Finding more links
- Including/excluding URLs

Starting points: Pages and Links

- **Base URL**
 - Used most commonly
 - `http://www.acme.com/`
- **URL File**
 - `C:\MyUrls.txt`
- **URL URL**
 - `http://intra.acme.com/urls.txt`
- Descendants indexed

Starting points: Pages only

- **Single Page**
- **Page File**
- **Page URL**
- Descendants not indexed

Finding more links

Documents other engines miss:

- JavaScript
 - **Execute JavaScript**
 - **Fetch JavaScript**
- Anytotx plugin
 - PDF links
 - Shockwave Flash links

Including/excluding URLs

- Expand or limit what gets crawled
- Filters, not sources:
 - URL must still be linked (direct/indirect) from Starting Points

Including/excluding URLs

Each URL must pass all of these tests:

- Any approved host or domain
- All required settings
- No exclusions

Any approved host or domain

Any of the following:

- Same site as any **Base URL**, **URL URL** or **URL File**
- Site matches any **Enterprise** or **Extra Domains**
- IP matches any **Extra Networks**
- URL matches any **Extra URLs REX**

Any approved host or domain

Enterprise or Extra Domains

- Domain suffix strings
- For multi-host crawls within a domain
- Example: **acme.com** allows:
 - **www.acme.com**
 - **shop.acme.com**
 - **support.acme.com**

Any approved host or domain

Extra Networks

- IP address prefixes
 - **10.10.** to cover 10.10.*.* intranet
- Useful for intranets with no-domain links:
 - **http://bob/**
 - **http://mary/**
- May slow crawl: more hostname lookups

Any approved host or domain

Extra URLs REX

- REX expressions to match URLs
- For additional host but only certain URLs
- Example:
 - **>>http://archives.acme.com/!widget*widget=**
 - Matches anything on **archives.acme.com** with **widget** in URL
- Negates **Stay Under** if matches top

Including/excluding URLs

Each URL must pass all of these tests:

- Any approved host or domain
 - **Base URL** **URL URL** **URL File**
 - **Enterprise** **Extra Domains**
 - **Extra Networks**
 - **Extra URLs REX**
- All required settings
- No exclusions

All required settings

URL must match all of these tests:

- Any **Extensions** (or **All Extensions = Y**)
- Any **Required Prefix** or **Required REX**
- **Stay Under** (per-site)
- Any **Protocols**

All required settings

- **Extensions**
 - URL file extensions
 - `.html .htm .txt .pdf .asp`
 - All URLs must match one
 - Anchor/query string ignored first
 - Extensions match this URL:
 - `http://www.acme.com/site.asp?id=103`
- **All Extensions**: set to **Y** to match any

All required settings

- **Required Prefix**
 - URL prefix(es)
 - `http://www.acme.com/shopping/`
 - Must be under `www.acme.com/shopping` tree
- **Required REX**
 - URL REX expression(s)
 - `>>=http://www.acme.com/!=catapult*catapult=`
 - Must be on `www.acme.com` and contain `catapult` in URL

All required settings

- **Stay Under**
 - **Y**: URLs must be under a Base URL
 - Negated by **Extra URLs REX** (more specific)
- **Protocols**
 - HTTP, HTTPS, FTP, Gopher, File
 - Also applies to embedded links (frames)

Including/excluding URLs

Each URL must pass all of these tests:

- Any approved host or domain
- All required settings
 - **Extensions** or **All Extensions**
 - **Required Prefix** or **Required REX**
 - **Stay Under**
 - **Protocols**
- No exclusions

No exclusions

URL must *not* match any of these:

- Any **Robots**
- Any **Exclusions**
- Any **Exclusion Prefix**
- Any **Exclusion REX**
- Any **Exclude by Field**
- **Off-site Pages**

No exclusions

Robots

- Allows *site* admin to exclude URLs
- Knows more about site than crawler admin
- Can protect load-sensitive parts of site
- Usually on

No exclusions

- **robots.txt**
 - If \mathcal{Y} , /**robots.txt** fetched from each site
 - Excludes by URL prefix and/or User-Agent
- **Meta**
 - If \mathcal{Y} , respect **<META>** robots tags in individual pages
 - More flexibility for site admin
 - **NOINDEX** vs. **NOFOLLOW** options
 - **<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">**

No exclusions

Exclusions

- URLs must not have any of these substrings
- Wildcards allowed: *
 - Not usually needed: already taken as substring
- Example:
 - **Exclusions** set to ~
 - Excludes **http://www.acme.com/~bob/**

No exclusions

- **Exclusion Prefix**
 - URLs must not have any of these prefixes
 - **http://www.acme.com/personal/**
- **Exclusion REX**
 - URLs must not match any of these expressions
 - Example: **[^\a\lnum]test[^\a\lnum]**
 - Excludes URLs containing **test**
 - Does *not* exclude **contestant**, **fastest**, etc.

No exclusions

Exclude by Field

- For complex exclusions
- Full Metamorph query syntax
- Check fields other than URL (Body etc.)
- Exclude pages vs. links
- Slower than other methods
 - Page must be fetched
 - More processing

No exclusions

Exclude by Field

- **Query**
 - Metamorph query: exclude if field matches
- **Meta or Field**
 - What field to search
 - **<META>** field searched if **Meta** specified
 - Otherwise **Field** searched
- **Exclude**
 - How to exclude on match
 - Pages and links: discard matching page and its links
 - Pages only: discard matching page but keep its links
 - Links only: keep matching page but discard its links

No exclusions

Off-site Pages

- If \mathcal{Y} , off-site pages are ok
- For one-off pages of other sites
- Links not walked (ala Single Page)
- **Multiple Fetches**
 - Rarely needed
 - Decreases crawl performance

Including/excluding URLs

Each URL must pass all of these tests:

- Any approved host or domain
- All required settings
- No exclusions
 - **Robots**
 - **Exclusions**
 - **Exclusion Prefix**
 - **Exclusion REX**
 - **Exclude by Field**
 - **Off-site Pages**

What to crawl: Review

- Starting points
 - **Base URL** etc.
- Finding more links
 - **Execute JavaScript**
- Including/excluding URLs: pass these tests
 - Any approved host or domain: **Extra Domains** etc.
 - All required settings: **Required Prefix** etc.
 - No exclusions: **Exclusion Prefix** etc.

How to crawl: Rewalk Type

- **New**
 - New database created each walk
 - Start from scratch at **Base URL**
 - Replaces live search only when done
- **Refresh**
 - One database for search and crawl
 - Pages refreshed in-place, instantly visible
 - Often-changing pages refreshed more often

Rewalk Type: New

- Pros:
 - Most accurately reflects current site
 - Live search database undisturbed and faster
- Cons:
 - More bandwidth and temp disk space used
 - Site changes take longer to show up

Rewalk Type: Refresh

- Pros:
 - Faster crawls
 - Site changes show up much faster
 - Less bandwidth and temp disk space
- Cons:
 - May get out of sync (rarely) with site
 - Search may be slowed if many site changes
 - Requires **If-Modified-Since** site support

When to crawl

- Manually
 - Hit **Go** or **Update and GO**
- Automatically: **Rewalk Schedule**
 - At specific time: specify
 - At site change: **Watch URL**

When to crawl

- **Notify**
 - Mail this address at end of scheduled walk
- **Attach Logs**
 - What to attach to mail
 - Walk Errors
 - Walk Status
 - Vortex Log
 - Monitor Log

When to crawl

- Refresh picks next time to fetch each page
- Based on page modification
- Can be tuned
 - **Default Refresh Time**
 - **Minimum Refresh Time**
 - **Maximum Refresh Time**

Data manipulation

- More data
- Changing HTML
- Changing search text
- Splitting large pages
- Duplicate prevention
- Alternate data sources

More data

Meta Data

- **Meta Tags**
 - Other specific **<META NAME>** tags
- **Standard Meta**
 - Keywords and description
- **All Meta**
 - All **<META NAME>** data

More data

JavaScript

- **Execute JavaScript**
- **Fetch JavaScript**
- May get body data as well as links
 - `document.write()`

More data

- Anytotx plugin
- Extracts text from PDFs, Word, Excel etc.
- May need to add to **Extensions**

Changing HTML

Keep HTML

- **Alt Text**
- **<STRIKE>**
- ****
- Form: **<SELECT>** **<INPUT>**
<TEXTAREA>

Changing HTML

- **Ignore Tags**
 - Remove HTML between these substrings
 - Example:
– `<!--BEGINIGNORE-->` `<!--ENDIGNORE-->`
- **Keep Tags**
 - Remove HTML *outside* these substrings
 - Example:
– `<!--BEGINCONTENT-->` `<!--ENDCONTENT-->`

Changing search text

- **Remove Common**
 - Removes header/footer common to all pages
 - Useful for removing navigation text
- **Ignore Characters**
 - Characters to strip from text and queries
 - e.g. – (hyphen) for part numbers

Splitting large pages

Plugin Split

- Split anytotx output into multiple pages
 - e.g. PDFs by page, ZIPs by file or depth
- Each result shares same URL
- Anchor contains file or page number

Plugin Split

- **Depth**
 - For splitting file archives (e.g. ZIPs)
 - What depth to split at (0: do not split by depth)
 - e.g. 1 to split ZIP, 2 to split ZIPs within ZIP, etc.
- **Bytes**
 - For splitting large files with no boundaries
 - How many bytes for each part (0: do not split by bytes)

Plugin Split

- **AtPage**
 - Whether to force Bytes splitting at a page boundary
 - Cleaner splits, but could be 50% larger than **Bytes** limit
- **Pages**
 - For splitting page-delimited files (e.g. PDFs)
 - How many pages per part (0: do not split by page)
- If both **Bytes** and **Pages** enabled:
 - Split at first limit reached
 - e.g. page-bounded PDFs and large unpaginated docs

Duplicate prevention

Prevent Duplicates

- Same-search-text documents stored once
- Useful if multiple URLs have same content
- Uses hash code (fast but not infallible)

Duplicate prevention

Duplicate Check Fields

- Concatenation of these fields is checked
- Default is just **Body**
 - Primary search field
 - Two same-**Body** docs considered duplicates, even with different Keywords

Alternate data sources

Data from Field

- Alternate sources for:
 - **Modified Date**
 - **Title**
 - **Description**
- Useful if If-Modified-Since unsupported

Data from Field

- **Search:** Metamorph query to match data
- **Replace:** Optional replacement for match
- **Meta or Field:** What field to search
 - **<META>** field searched if **Meta** specified
 - Otherwise specified **Field** searched
- **Which Field:** where to save match data
 - **Modify Date**
 - **Title**
 - **Description**

Resource limits

- Limit resources used (memory, CPU etc.)
- Saves load on crawler
 - Usually same as search machine
- Saves load on targeted site
- Avoid bogging down on large/slow sites

Resource limits

- **Crawl Delay**
 - Saves load on site
 - Disables threads
- **Parallelism**
 - **Threads:** per process (single site)
 - **Servers:** simultaneous processes (1 per site)
 - Increases load on crawler
 - May not increase crawl speed

Resource limits

Per-page limits

- **Max Page Size**
 - Increase for large source pages (PDFs)
- **Page Timeout**
 - Increase for slow sites
- **Max Frames**
 - 0 to treat as separate pages
- **Max Redirects**
 - 0 allows none
 - NTLM (Windows) auth counted as redirect

Resource limits

Overall crawl limits

- **Max Pages**
- **Max Bytes**
- **Max Depth**
 - Depth is number of links from **Base URL**

Resource limits

Crawl pauses:

- **Maximum Process Size**
- **Maximum Load Average**
- Measured by crawler, not web site
- Exceeding either pauses walk
- Continues at next crawl

Authorization

- For crawler (Search Settings covers search)
- Should provide credentials to access entire site
- **Login Info**
 - **Name**, **Password** for protected sites
 - Basic, NTLM, FTP, Windows file://
- If firewall/proxy access needed:
 - **Proxy**
 - **Proxy Login Info**

Authorization

- **Primer URL**
 - For cookie-protected sites
 - Usually site login page: gets **Login Info**
 - Fetched at start of each process
- **Cookie Source Path**
 - Pre-load cookies from file (e.g. browser save)
 - Usually **Primer URL** with **Login Info** suffices

Miscellaneous crawl settings

- Debug messages
- Fixing rare errors
- Headers
- Links

Debug messages

- **Notes**
 - Quick-ref notes for admins
 - Not used by crawler
- **Verbosity**
 - 0 Show errors
 - 1 Also show starting point URLs
 - 2 Also show some setting info
 - 3 Also show URLs found in URL files
 - 4 Also show why URLs rejected
- **Debug JavaScript**

Fixing rare errors

Typically changed by tech support advice

- **Entropy Source**
- **DNS Mode**
- **Net Mode** (Webinator only)
- **Temporary Dir** (Webinator only)
- **Max Requests**

Headers

- **User Agent**
 - Sets User-Agent header
 - For sites that "don't support" certain browsers
- **Mime Types**
 - Sets Accept header
 - Site may return better version of content
 - Up to site to respect it

Links

- **Inline Iframes**
 - **Y**: Fetch **<IFRAME>**s and embed in containing page
 - **N**: Treat as separate links
- **Max Frames**
 - Max **<FRAME>**s/**<IFRAME>**s to fetch per page
 - 0: Treat frames as separate links

Links

Embedded Security

- For embedded objects only (frames, scripts)
- **Non-decreasing**: no http objects in https page
- **Non-increasing**: no https objects in http page
- **same-protocol**: protocol must match page (e.g. no file objects from HTTP page)

Links

- **Strip Queries**
 - Removes ? and beyond from URLs
 - Rarely used: query often shows page content
- **Ignore Case**
 - URL path case ignored (lowercased)
 - For sites with mixed-case links (non-standard)

Links

- **Index Name**
 - What files to consider same as dir in URL
 - Default `index.html` `index.htm`
- **Store Refs**
 - Whether to remember page relationships (parent/child)
 - If set to **N**:
 - Saves some time and disk space
 - Prevents Children/parents reporting in List/Edit URLs
 - Less effective error reporting

Search settings

- Appearance Options
- Query Options
- What to search
- Logging
- Spell checking

Appearance Options

- **Result Order**
- **Result Style**
- **XSL File** (Appliance only)
- **Abstract Style**
- **Abstract Length**
- **Results per Page**
- **Results Width**
- **Box Color**
- **Display Thunderstone logo** (Appliance only)

Appearance Options

- **Show Advanced Search**
- **PDF Query Highlighting**
- **Font**
- **Top HTML**
- **Bottom HTML**
- **Enable Sherlock**
- **Apply Appearance**
- **Revert Appearance**

Query Options

- Query syntax
 - What syntax user may use
- Query processing
 - How query is processed
- Relevance ranking
 - Ranking and ordering results

Query syntax options

Thesaurus options

- **Synonyms**
 - Disabled
 - Phrase recognition only
 - Phrases & Allow synonyms
 - Phrases & Use synonyms by default
- **Main Thesaurus**
- **Secondary Thesaurus**

Query syntax options

- **Allow the @ Operator**
- **Allow Linear**
- **Allow "NOT" Logic**
- **Allow Post-Processing**
- **Allow Wildcards**
- **Allow "WITHIN" Operators**
 - Overrides **Proximity**
 - Requires post-processing if not **w/page**

Query processing options

- **Resolve Phrase Noise Words**
- **Keep Noise Words**
- **Fast Result Counts**
- **Proximity**
 - **Line, Sentence, Paragraph, Page**
 - Requires post-processing if not **Page**
- **Word Forms**
 - **Exact match**
 - **Plurals & possessives**
 - **Any word forms**

Relevance ranking options

Rank knobs: off/low/medium/high/max

- **Word Ordering**
- **Word Proximity**
- **Database Frequency**
- **Document Frequency**
- **Position in Text**

Relevance ranking options

Ranked Rows

- Max results from relevance-ranked query
- Default 200
- Larger is slower, users may not page that far
- Summary still shows hit counts over this

What to search

- Under All Walk Settings because index affected
- Any change causes index re-build (may take time)
- **Word Definition**
 - REX expression(s) to define words
 - Default:
 - `[\a1num\x80-\xff]{1,70}`
 - `[\a1num\x80-\xff.]{1,70}`
 - `>>[.&']=[\a1num\x80-\xff.]{1,70}`

What to search

Index Fields

- Under All Walk Settings actually
- What fields to search for user's query
 - Any combo of **Title, Description, Keywords, Meta, Body**
- Field order affects ranking
 - If **Position in Text** is on

Logging

- **Query Logging**
 - Whether to log users' queries
 - Top queries, no-hits, best-bet clicks reports
- **Rotate Schedule**
 - When to rotate (clear) query log
- **Email**
 - Nonempty: Also email logs to this address

Spell checking

- **Enable Spell Check**
 - Whether to suggest “Did you mean” searches
 - Suggestions based on actual content
- **Suggest Time Limit**
- **Number of Suggestions**

Best Bets

- Show pre-selected links adjacent to results
- Based on keywords in user's query
 - Suggest useful links
 - Promote/advertise URLs
- Config:
 - Define group
 - Add URLs to group
 - Enable group for profile

Best Bets

Define group: **Best Bets Groups** menu item

- **Group Name**
 - For later reference
- **Result Type**
 - Combo of **Title**, **Description**, **URL** to show users

Best Bets

Add URLs to group: **List/Edit URLs** menu

- Search and choose URL to add
- **Best bet words**
 - **Priority** (highest shown first)
 - **Title** (can differ from walked title)
 - **Keywords** (user query must match)
 - **Group** that this URL belongs to
 - **Description** to also show user

Best Bets

Enable group for profile: **Search Settings**

- Top and/or Right group (relative to results)
- **Title**, **Group**, **Color**, **Style** for each
- At search:
 - User's query run against **Group's Keywords**
 - Matching URLs are displayed

Categorization

- Add category-specific user search
 - e.g. Sales, Support, Reference
- List appears in select box on search form
- Search Everything or a particular category
- Config under **All Walk Settings**
 - Managed at crawl time

Categorization

- **Category:** User-visible name (e.g. Sales)
- **URL Pattern:** Full URL pattern
 - Wildcards: * or ?
 - Not a substring: need trailing * unless single-page
 - e.g. `http://www.acme.com/sales/*`
- **Visible:** whether to show on search form
 - Make search select box more concise
 - Still searchable (e.g. custom static form)
- Changing settings forces re-categorization

Character set issues

- Settings
- Look and feel
- Document types

Charsets: Settings

- **Storage Charset**
 - Data is stored and searched as this charset
 - Typically UTF-8
 - Most browsers query in UTF-8
 - Little or no character loss (PDFs etc.)
- **Source Default Charset**
 - Rarely changed (e.g. unlabelled docs)

Charsets: Settings

- **Display Charset**
 - What charset to display results in
 - Should agree with look and feel
 - Most efficient if same as **Storage Charset**
- **Top HTML**
 - Should set charset via **<META>** tag
 - Tag updated to reflect **Display Charset**

Charsets: Settings

- **XML UTF-8**
 - Whether to make data XML-safe at crawl (vs. search time)
 - **N:** supports everything
 - **Y:** supports XML, not all hi-bit chars, no PDF highlighting
 - Meaning changed in version 5.4.7

Appliance Features

- Query via Program: XML output
- Meta Search
- Results Authorization
- Crawling Filesystems
- Custom Thesaurus
- Replication
- Access Control
- System Maintenance

Query via Program

- Issue POST or GET to:
 - Appliance:
 - `http://host/texis/search/`
 - Webinator:
 - `http://host/texis/webinator/search/`
- Search variables needed (others in manual):
 - **pr**: profile to search
 - **query**: search terms

Query via Program: XML output

- **Allow XML** set to **Y**
 - Permits XML results regardless of **Result Style**
- Query `/texis/search/main.xml`
- Same search variables
- Results returned in XML
 - Variables described in manual

Meta Search

- Search multiple profiles as if one (meta)
- No effect on walks
- Typical uses:
 - Walk with multiple profiles but search as one
 - Faster search of large data (multiple machines)
- Back-end profile(s) config
- Meta (front-end) profile config

Meta Search: Back-end profile

- Search Settings: **Visible**
 - Allows profile to be searched by meta searches
- **Maintenance / System Wide Settings / Cluster Members**
 - "Friends" to accept data or searches from
 - Exact IP address: `127.0.0.1`, or
 - IP prefix and * wildcard: `10.10.10.*`

Meta Search: Meta profile

- Meta profiles are special
- Create as copy of `-Meta Search Defaults-`
- Search only (no walk)
- Walk settings replaced with Meta Search

Meta Search: Meta profile

- **Profiles**
 - List of back-end profiles to search
 - **Host IP or Name:** IP (preferred) or host
 - **Profile Name:** name of back-end profile
- **Visible**
 - Allows profile to be searched by other metas
 - “Daisy-chain” meta searches

Meta Search: Review

- Back-end config
 - Set profile **Visible**
 - Add meta search hosts to **Cluster Members**
- Meta (front-end) config
 - Create as copy of **-Meta Search Defaults-**
 - Add back-end profiles to **Meta Search Settings**

Results Authorization

- Limits search results by document
- Only show what user can see
- Each search result verified at search
- Controlled by existing security (**Authorization Method**):
 - Cookies
 - Basic/NTLM/File

Results Authorization

- **Forward login cookies**
 - For cookie-based server security
 - Appliance forwards cookies at verify
- **Login Cookies**
 - List cookie(s) that control security
- **Login URL**
 - Redirect for user if missing cookies

Results Authorization

- **Basic/NTLM/file - prompt via form**
 - For Basic/NTLM/file protected documents
 - Appliance will prompt for user/pass
- **Basic/NTLM/file Cookie Type**
 - For saving user/pass
- **Login Verification URL**
 - User/pass checked against this
 - Should require login, but *any* login

Results Authorization

- **Unauthorized Result Query**
 - If server denies with 200 Ok instead of 401 Unauthorized
 - **Field/Type:** What to search
 - Text/substring
 - Text/REX
 - HTML/REX
 - **Query:** What to search for

Crawling Filesystems

- Windows or Unix filesystems
- Mount filesystem
 - Two methods (new/old)
- Crawl it

Mounting: New method

- Version 5.4.11 or later
- **Maintenance / Network Filesystems & Shares**
- Shows current mounts
- Add: **NFS - Unix** or **SMB - Windows**

Mounting: New method

NFS - Unix

- **Server:** hostname (case insensitive)
- **Directory:** full path as exported (case sensitive)
 - e.g. `/documents/internal`
- **Reliability:**
 - **Hard:** retry
 - **Soft:** give up, continue elsewhere
- **NFS Version:**
 - 3 unless problems with old servers

Mounting: New method

SMB - Windows

- **Server:** hostname (case insensitive)
- **Share:** as exported
 - e.g. `internal`
- **Login Name**
 - To login on Server
 - Needs access to all files
- **Login Password**
- **Server IP**
 - Rarely used: Only if DNS name differs from Server

Mounting: Old method

- **Maintenance / Webmin Interface**
- **Disk and Network Filesystems**
- **Add mount** (NFS or Windows)

Crawling Filesystems

- **Base URL**
 - `file://corpserv/internal/...`
- Users' browsers must access the same way

Custom Thesaurus

- Replace or add terms to standard thesaurus
 - Industry-specific terms
- Edit locally
- Upload
- Profile use

Custom Thesaurus: Editing

- “User equivalence file format” in Taxis manual
- One per line
- Root word followed by equivs
 - **ranch, farm, hen house, pen**

Custom Thesaurus: Upload

- **Maintenance / Custom Thesaurus**
- **Name:** symbolic name to refer by
- **Permutations:**
 - **None:** root finds root or equivs
 - **Single:** and equiv finds root
 - **Full:** and equivs find equivs
- **New file:** local source file (save it)
- **Uses:** number of profiles using it

Custom Thesaurus: Profile use

- Custom Thesaurus selectable for:
 - **Main thesaurus**
 - **Secondary thesaurus**
- Affects:
 - **Synonyms**

Replication

- Automatically copy data to another profile
- Typical uses:
 - Combine multiple profiles into one for search
 - Back up walk data to separate machine(s)
 - Load balance search across multiple machines
- Sender config
- Receiver config

Replication: Sender config

- **Replication Settings**
- List of receiving profiles
- **Host IP or Name**
 - Hostname of receiver (e.g. **localhost**)
- **Profile Name**
 - Name of receiving profile (can be created after)

Replication: Sender config

- **Update and GO**
- **Walk Status**
 - Pages walked added to Replication Queue
 - Queue to be sent to Receiver

Replication: Receiver config

- Need to accept Sender
- **Maintenance / System Wide Settings / Cluster Members**
 - "Friends" to accept data or searches from
 - Exact IP address: **127.0.0.1**, or
 - IP prefix and * wildcard: **10.10.10.***

Replication: Walk Status

- **Walk Status** for Sender profile:
 - Replication queue decreases as items sent
- **Walk Status** for Receiver profile:
 - Shows pages "walked"
 - Actually received from Sender

Replication: Review

- Sender config
 - Add Receiver(s) to **Replication Settings**
 - Start walk
- Receiver config
 - Add Sender(s) to **Cluster Members**

Access Control

- Finer control over admin (not search) access
- Typical uses:
 - One-profile admins
 - Look-and-feel admins
- Concepts
- Config
- What rights needed when

Access Control: Concepts

- User groups
- Object hierarchy
- Access Control Lists
- Determining effective rights

Access Control: User groups

- Simplify admin user maintenance
- Contains users and/or other groups
- User inherits perms via membership
 - User **Amy** is in group **Programmers**
 - **Programmers** is in group **IT**
 - **Amy** has **Amy**, **Programmers** and **IT** perms
- Special group **Everyone**

Access Control: Object hierarchy

- Object is access-controllable action or thing
- Some are supersets:
 - Edit all profiles vs. specific one
- Hierarchy arranges supersets
- Object inherits perms from ancestors
 - If user can edit all profiles, user can edit specific one

Access Control: Object hierarchy

```
/
  Users/
    admin
  ...
  Groups/
  Profiles/
    default
  Settings/
  Maintenance/
```

Access Control Lists

- Determines what rights users have on an object
- Each object may have one
- Composed of Access Control Entries (ACEs)
- An ACE contains:
 - Trustee (user or group)
 - List of rights
 - Allow or Deny

Determining effective rights

- Use first ACE matching both user and right
- ACEs examined in specific order
 - First: ACEs explicitly set on object
 - Then: ACEs set on ancestors (nearest first)
- Each object's ACEs checked in ACL order
- If no match, access allowed by default

Access Control: Config

- Enable/disable all access control
- **User Groups** menu
- **Access Control** menu

Config: Enable/disable

- Only **admin** user can enable/disable
- **Maintenance** menu
- **Enable Access Control Lists**
 - Adds Access Control, User Groups menus
- **Disable Access Control Lists**
 - Disables access control
 - Deletes all user groups and ACLs

Config: **User Groups** menu

- Shows all groups and their members
- **Add a Group**
- **Edit**
 - Add/remove members
- **Delete**

Config: **Access Control** menu

- Shows all ACLs
- **Add an ACE**
 - Which sub-object (or all)
 - **Trustee**: user or group
 - **Type**: **Allow** or **Deny**
 - **Perms**: Read/Write/Delete/Change Perms
- ACE order is important

What rights needed when

- **Profiles/** and **Settings/** are siblings
- Need both (intersection on grid)
- Allows flexible user rights
 - One profile, all settings
 - One setting, all profiles

What rights needed when

Walk / Search Settings

- To see: Profile read and setting read
- To change: Profile write and setting write+delete
- If no read access: setting not shown at all
- If no write access: setting grayed out

What rights needed when

- Start/stop walks
 - Profile write and **Walk now** setting write
- **Best Bets** edit (groups/URLs)
 - Profile write and **Best Bets Groups** setting write
- **List/Edit URLs**: edit URLs / **Update Soon**
 - Profile write and **Link report** setting write

Access Control: Review

- User groups: Easier maintenance
- Object hierarchy: Actions/things to control
- ACLs: Determine rights
- Inheritance from groups/ancestors

Access Control: Example

- Initial lockdown (in this order)
 - Allow **admin** all rights Global
 - Deny **Everyone** all rights Global
- Locks out all but **admin** (e.g. new users)
- Enable other users as needed

Access Control: Example

User **Bob** fully controls profile **Archives**:

- Lockdown as above
- Profile ACE on **Archives**
 - User **Bob**, read and write, **Allow**
- Setting ACE on **All Settings**:
 - User **Bob**, read write and delete, **Allow**

System Maintenance

- **Information**
- **Install/Upgrade**
- **Logs**
- **Search Appliance Settings**
- **Appliance system access**

Maintenance: Information

- **Display disk space**
- **System Information**
- **Thunderstone Information**
- **Tech Support Information**

Maintenance: Install/Upgrade

- **Setup/edit update preferences**
 - Automatically install software updates
- **Check for updates**
 - Look for software updates now
- **Install from CD**
 - Install/update behind firewall

Maintenance: Logs

- **View Taxis logs**
 - `error.log`, `monitor.log`,
`transfer.log`, `vortex.log`
- **Send Taxis logs to Thunderstone**
 - `boot cron messages`
- **View System logs**
- **Send System logs to Thunderstone**

Maint.: Search Appliance Settings

- **System Wide Settings**
 - Customize home page
 - Default profile
 - Set `favicon.ico`
 - **Cluster Members**
- **Save/Restore Search Appliance settings**
- Others discussed previously

Maint.: Appliance system access

- **RAID Array Management**
 - New for certain hardware configs
- **Webmin Interface**
 - Network, disk, server, time config
 - Usually pre-set