IAFA Templates in Use as Internet Metadata

David Beckett, Computing Laboratory, University of Kent, Canterbury, CT2 7NF, England
`D.J.Beckett@ukc.ac.uk, http://www.hensa.ac.uk/parallel/www/djb1.html`

Abstract:

Recently there has been a growing need for a metadata standard for the Internet. The files that are available on ftp and WWW sites can be difficult to search if they are enclosed in a container format (e.g. tar). and bibliographical data can be deeply embedded in documentation. This paper describes how IAFA Templates[Deutsch95] have been used in a real archive to store the metadata of lots of different types of documents and software and to derive WWW, gopher and text indices from them.

Keywords:

Metadata IAFA Templates ALIWEB SOIF Harvest

Introduction

Despite the popularity of the HTTP/HTML part of the web, the most standard form for transmitting and sharing documents and software on the internet is still via ftp and gopher sites. These have grown to be large resources of material, but unfortunately have been traditionally very badly indexed and organised.

If a gopher interface to an archive is available it can provide a menu based interface in which the archive administrator can describe the resource, albeit in around a maximum of 70 characters for common terminal sizes, which is what most common gopher clients run on.

More commonly, only the anonymous `ftp` form was available, providing just the UNIX shell-like interface to the archive.

Usually, the donated files were lucky to have a single line of text describing the contents, or more likely, the filename was the best hint to the package (`foobar-1.3.tar.gz` for version 1.3 of package *foobar*). Sometimes, there would be a `README` or index file in the same directory as the files with a description of each of the files in natural language. This would be fine for people prepared to look at every single `README` file in the archive to find something but natural language is not a good way to describe the files since the information is not structured and hence not machine readable / writable. There are also additional problems:

Difficult to indexing locally

The text in the `README` files could be fully indexed (inverted) but there would be no way of picking out the descriptions for individual files from the text.

Cannot do resource discovery well

For similar reasons to above, remote indexers would have difficulty picking out files and would have to index the content of the files without really knowing what is there.

Index sharing is not possible

There is no way to share indices without having a standard index / metadata format.

Hence the need for structured metadata standards for Internet archives.

Metadata standards for the Internet

In May 1993 I started to build a Parallel Computing Archive[1] at the HENSA Unix Archive[2]. The materials gathered consist of software, papers, reports, bibliographies, documents and many other types of file, taken from several sources:

- Locally written.
- Donated from external contributors.
- From off-line sources.
- *mirrored*[McLoughlin95] from other Internet sites.
- Automatically archived such as USENET newsgroups.

In the first three cases, it would be easy to only allow materials with correctly formatted metadata to be allowed on the archive, but the the latter cases are more difficult. *mirroring* is a process which makes an identical copy (a clone, or mirror image) of the files on a remote site on the local site. Thus, these files cannot be modified locally, and any metadata must be external, in other files. For the newsgroups, there are a lot of articles being archived daily so generating appropriate metadata by hand would be very tedious. Thus to handle all of the above sources, a metadata standard was needed with the following requirements:

- Easy for people to read and write.
- Machine readable and writable for automatic creation, modification and indexing and sorting.
- Can describe the *form*, *contents* and *location* of the information.
- Structured to allow nesting.
- Can be used for building multiple derived indices (WWW, text, gopher, ...)

Investigations were done into the metadata forms that were available at the time:

Linux Software Map (LSM) templates

The Linux software archives at the SunSITEs[3] addressed their need for a metadata standard with structured *templates*[Kopmanis94] which contain the following 12 attributes appropriate to the archive needs:

> Title, Version, Entered-date, Description, Keywords, Author, Maintained-by, Primary-site, Alternate-site, Original-site, Platforms, Copying-policy

The form of the entries is similar to Internet Mail headers[Crocker92] with colon separated attribute-value pairs that can wrap over several lines. There is a short description of the valid values for each field but little concrete data form, most of it is free form text.

Later on, tools were built to process these templates, index them and create such things as the *Linux Software Map*[Boutell95]. At present, when people are submitting something to the archive, they may be rejected by the maintainers if they do not have LSM templates written.

Unfortunately, the LSM templates are very much intended for software packages that are replicated at different sites and hence are not particularly appropriate for indexing a much richer set of files.

IAFA templates

The Internet Engineering Task Force (IETF) Working Group on Internet Anonymous FTP Archives (IAFA), later called IIIR, have produced the IAFA templates Internet Draft[Deutsch95]. This defines a range of indexing information that can be used to describe the contents and services provided by anonymous FTP archives.

The draft has a rich range of templates, attributes and values that can be used to describe common and useful elements. The goal is that these are to be used to index archives, made available publicly in them to allow searching, indexing and sharing of information on the archive contents, services and administrative data.

This template scheme is based on the same RFC822 form like the LSM templates, with colon separated attributes-values known as data elements. One or more data elements are collected into templates which have a single `Template-Type` field to describe the type of the basic template. Multiple templates can be collected in index files by separating with blank lines. The attributes can be structured in several ways:

- *Variant* information which are used to support multiple languages, formats, ... of a document, for example: `Language-v0: English` and `Language-v1: Deutsch` describe two variants of language available for an individual resource.
- *Clusters* which are classes of data elements which occur every type an individual or group is mentioned such as name, addresses, email addresses, telephone numbers etc. Handles can be used to refer to clusters inside templates.
- *Handles* which allow short unique strings to abbreviate a group of data elements for individuals or organisations. For example, `Author-Handle: Kim Jones` instead of all the individual elements of the USER cluster for Kim Jones.

There are 14 currently defined template types:

> SITEINFO, LARCHIVE, MIRROR, USER, ORGANIZATION, SERVICE, DOCUMENT, IMAGE, SOFTWARE, MAILARCHIVE, USENET, SOUND, VIDEO, FAQ

and each has appropriate attributes defined for them. Most of the types are self explanatory apart from SITEINFO which is a description of the FTP site and LARCHIVE which is a description of a logical (sub-)archive. More types can be defined if necessary having the same basic attributes as DOCUMENT.

It also turns out that LSM Templates were based on an early draft of the IAFA Templates (June 1992) but modified to be have more consistent elements. The later versions were modified to be more similar but some differences remain.

IAFA Templates were the solution chosen to base my metadata on. They were rich and extensible and a standard, or albeit a draft one.

Using IAFA Templates

The first stage was to convert all the old `Index` files that had been written by hand, into IAFA Template form. This was achieved by just mapping path, description pairs for each file into a simple form:

```
Template-Type: DOCUMENT
URI: path
Description: description
```

but of course, not everything is a document and more intelligence was needed to determine the metadata.

Extensions to IAFA templates

IAFA templates were not sufficient to fully handle all my uses so new template types and elements were added, as the draft allows.

The information is structure hierarchically and hence there is a need to list the sub-directories for any given directory. There is no way to do that cleanly in the draft, the only way would be to rely on a convention that a `DOCUMENT` with a URI ending in a '/' is a directory. It is better to add a `Template-Type DIRECTORY` since a directory is **not** a document. Another template type that was added was `EVENT` which was used with to describe conferences, workshops, etc. which have a date range.

In addition, there was no way to describe symbolic links. These are used in my archive to point from one area to another so that the directories /parallel/transputer/compilers/occam and /parallel/occam/compilers have the same content but the names of the final directories are different. If the alternative was used, a site-relative URI, that would make the directory names the same and hence confusing for the browser. A simple extension to the format of the URI field allowed symbolic links to be added.

Another extension was the definition of the separation of templates. The draft uses a *blank line* defined as an empty line or a line consisting of only white space. I use just the former, an empty line since that means paragraph breaks can be put in descriptive or other text (see example later).

Extra elements that were added include:

`X-Abstract`
: The abstract from a paper or report. This is more specific than a general `Description` field.

`X-Acronym`
: Rather than put the (say, conference) acronym in `Title`, it goes here.

`X-Gopher-Description(-v*) X-HTML-Description(-v*)`
: Descriptions that are specifically written for gopher or HTML index output. Gopher ones need to be short to fit on the screen and HTML ones can have markup added.

`X-Start-Date X-End-Date`
: For documents that describe a range e.g. conferences

`X-Expires-Date`
: Documents that can be deleted at a certain point for example job offers and conference calls.

Implementation of IAFA Templates

This was written in Perl 5 as two programs. The first one, `update-afa-indices` updates the IAFA indices: deleting templates for files that have gone; updating them for files that have changed in size and/or date and adding new templates for new files. The second program, `afa-to-others` reads the IAFA indices and outputs several derived indices: text, gopher and HTML.

Automatically updating IAFA Templates (`update-afa-indices`)

This program implements three forms of automatic extraction of metadata:

1. **From the URI (filename)**
   This is a very cheap operation since it requires no access to the file system or reading of the file contents. This is what is commonly done with HTTP servers to define the MIME types of the files being delivered. Things that can be interpreted from this include the `Template-Type` and the `Format` of the the file. For example, files ending in `.ps` are assumed to be format `postscript`, template `DOCUMENT`, files ending in .tar, .tgz, .taz, .tar.gz, .tar.Z were interpreted as the various forms of (compressed gzipped) tar `SOFTWARE` templates. For multiple levels of nesting, such things as `uuencoded compressed tar file` formats are possible.

   In later versions of the draft, this was changed to be the MIME[Borenstein93] type of the document but this would not be sufficient to describe the tar files above so the earlier version of the definition was kept. MIME types could be added easily.
2. **From the file system information**
   This information is usually kept in one place in the file system and can be found in a quick access. Things that can be interpreted from this include the `Size` of the file in bytes and the `Last-Revision-Date`.
3. **From the contents of the file**
   This is expensive to create and update since it requires expanding the presentation nesting. Essence[Hardy95a] does sophisticated work on the contents but this software limited itself to extracting author information (`Author-Name` and `Author-Email`) from USENET and mail files.

Because all of the above are generated by software, when they change, the software can update the metadata.

`update-afa-indices` operates on a directory tree (or subtree of it). In each directory, there is a single index file `AFA-INDEX` containing the templates for each of the files and directories in there. There is also a configuration file `.ixconfig` that allows specific files or directories to be excluded from the index. This allows mirrored areas to be kept the same as the remote site, but not all the files need to be shown. For example, if the entire contents of a text index file are represented in the index, there is no need to include a reference to that file.

The program walks the tree, reading the IAFA indices and looks for differences between the templates and the entries in the file system for items 1 and 2 from the list above. Only if there is a difference, is item 3 is calculated, since it is an expensive operation. If the entry is new, it is appended to the end of the index. Files or directories that have been deleted are automatically removed from the index. After all the processing is done, the index file is sorted by fields that are configurable by the `.ixconfig` file.

Adding hand-written Metadata to the IAFA indices

In addition to the above automatically added metadata, there is scope for the administrator to add lots more fields which are difficult for automatic software to pick out such as `Title` and `Description` - the main one which describes the contents. These are not checked or altered by the software when it does updates.

Building derived indices from IAFA indices (`afa-to-others`)

This program reads the IAFA indices files and generates derived indices, specific to particular access methods. For example, given the following template:

```
Template-Type: EVENT
Description: Call for papers for the Fifth International Conference on Parallel
        Computing (ParCo'95) being held from 19th-22nd September 1995 at
        International Conference Center, Gent, Belgium.

        Topics:
        Applications and Algorithms; Systems Software and Hardware.

        Deadlines: Abstracts: 31st January 1995; Notification: 15th April
        1995; Posters: 30th June 1995.

        See also <URL:http://www.elis.rug.ac.be/announce/parco95/cfp.html>
Author-Email: a.n.author@host.site.country
Author-Name: A. N. Author
Title: Fifth International Conference on Parallel Computing
X-Acronym: ParCo'95
X-End-Date: 1995-09-22
X-Expires-Date: 1995-09-22
X-Start-Date: 1995-09-19
Format-v0: ASCII document
Format-v1: PostScript document
Last-Revision-Date-v0: Wed, Jan 11 11:24:39 1995 GMT
Last-Revision-Date-v1: Wed, Sep 21 10:41:01 1994 GMT
Size-v0: 4516
Size-v1: 71330
URI-v0: parco95.ascii
URI-v1: parco95.ps
X-Gopher-Description-v0: 5th Int. Conference on Parallel Computing
        (ParCo'95) CFP (ASCII)
X-Gopher-Description-v1: 5th Int. Conference on Parallel Computing
        (ParCo'95) CFP (PS)
```

which describes a pair of files for a conference call. The derived text index output would be:

```
parco95.ascii
        "Fifth International Conference on Parallel Computing"
        Call for papers for the Fifth International Conference on Parallel
        Computing (ParCo'95) being held from 19th-22nd September 1995 at
        International Conference Center, Gent, Belgium.
        Topics: Applications and Algorithms; Systems Software and Hardware.
        Deadlines: Abstracts: 31st January 1995; Notification: 15th April
        1995; Posters: 30th June 1995.
        See also <URL:http://www.elis.rug.ac.be/announce/parco95/cfp.html>
        Author: A. N. Author <a.n.author@host.site.country>. [ASCII document]

parco95.ps
        "Fifth International Conference on Parallel Computing"
        Call for papers for the Fifth International Conference on Parallel
        Computing (ParCo'95) being held from 19th-22nd September 1995 at
        International Conference Center, Gent, Belgium.
        Topics: Applications and Algorithms; Systems Software and Hardware.
        Deadlines: Abstracts: 31st January 1995; Notification: 15th April
        1995; Posters: 30th June 1995.
        See also <URL:http://www.elis.rug.ac.be/announce/parco95/cfp.html>
```

```
        Author: A. N. Author <a.n.author@host.site.country>. [PostScript document]
```

and the derived gopher elements would be these entries in the gopher tree:

```
5th Int. Conference on Parallel Computing (ParCo'95) CFP (ASCII)
5th Int. Conference on Parallel Computing (ParCo'95) CFP (PS)
```

and the HTML element would be (as part of a conformant HTML 2.0 index file):

```
<DL>
  <DT><A NAME="parco95.ascii" HREF="parco95.ascii"><STRONG>Fifth Internati\
onal Conference on Parallel Computing (<EM>ParCo'95</EM>)</STRONG></A> [\
ASCII document] (4516 bytes)<BR>
  <DT><A NAME="parco95.ps" HREF="parco95.ps"><STRONG>Fifth International C\
onference on Parallel Computing (<EM>ParCo'95</EM>)</STRONG></A> [PostScri\
pt document] (71330 bytes)<BR>
  <DD>Call for papers for the Fifth International Conference on Parallel
Computing (ParCo'95) being held from 19th-22nd September 1995 at
International Conference Center, Gent, Belgium. <P>

<EM>Topics:</EM>
Applications and Algorithms; Systems Software and Hardware.<P>

<EM>Deadlines:</EM> Abstracts: 31st January 1995; Notification: 15th April
1995; Posters: 30th June 1995.<P>

See also <A HREF="http://www.elis.rug.ac.be/announce/parco95/cfp.html">htt\
p://www.elis.rug.ac.be/announce/parco95/cfp.html</A><P>

Author: A. N. Author (<EM>a.n.author@host.site.country</EM>).
</DL>
```

which looks like this when displayed formatted:

Fifth International Conference on Parallel Computing (*ParCo'95*) [ASCII document] (4516 bytes)
Fifth International Conference on Parallel Computing (*ParCo'95*) [PostScript document] (71330 bytes)

> Call for papers for the Fifth International Conference on Parallel Computing (ParCo'95) being held from 19th-22nd September 1995 at International Conference Center, Gent, Belgium.
>
> *Topics:* Applications and Algorithms; Systems Software and Hardware.
>
> *Deadlines:* Abstracts: 31st January 1995; Notification: 15th April 1995; Posters: 30th June 1995.
>
> See also http://www.elis.rug.ac.be/announce/parco95/cfp.html
>
> Author: A. N. Author (*a.n.author@host.site.country*).

Extra template types that were added for the derived indices were: `X-AFA-HEADER` and `X-AFA-FOOTER` which were some text to be placed at the start or end of a derived index.

This software is also configurable by the `.ixconfig` file and it allows hand-written indices e.g. the top `index.html` which is the *home page*, to be left untouched. In addition, some areas can be left without indices, for example, directories containing icons used in the HTML pages.

Problems With IAFA Templates

There are some problems with the IAFA templates as they currently stand. As described above, some extra elements were needed for my application, and indeed they could be added. More fundamentally, there is a problem with the structuring of the nesting of data. There is no way to describe a collection that, for example, contains multiple languages and multiple document types.

There are also the problems of encoding; there is no way to use binary data, non-ASCII characters, or indeed, blanks in descriptions as paragraphs (without the extension I used). Some of these problems have been addressed in other formats, and other metadata standards for different purposes are being designed which may provide a rich-enough structure to cope with these difficulties.

New Metadata Formats

Several new metadata formats have appeared more recently, albeit some still in unfinished or draft form.

Harvest Summary Object Interchange Format (SOIF) and Harvest

The SOIF Data format as described in [Hardy95b] used by Harvest[Bowman94] is based on IAFA Templates and BibTeX but has some extra features. Unlike the templates, it was designed to support streams of (possibly compressed) SOIF data between systems allowing additions, deletions and updates of the metadata. This is used by the Harvest system programs to communicate. SOIF also allows binary content in the values, by adding a length element to each value. There are not yet any *required* attributes defined by the standard although some are proposed. IAFA templates can be easily converted into SOIF format according to Koster in his *Future of ALIWEB* discussion[Koster94].

Universal Resource Citations (URCs)

The latest Internet Draft[Daniel95a] describes one of the main uses of a URC is to map from a URN to a possibly empty list of URLs for a browser. The user may, however, want to take the URC for the resource and find out the metadata of the URLs - cost, bibliographical data, etc. in a form that are understandable by people. The requirements for URC also include that it must be parsable by a computer, be simple and structured for nesting. Two URC services have been proposed in Internet Drafts, a simple text one in [Hoffman95] and one using SGML[Daniel95b].

Dublin Metadata Core element set

In March 1995, the OCLC/NCSA Metadata Workshop was held in Dublin, Ohio, USA with selected invited attendees from librarianship, computer science, text encoding, and related areas. One of its goals was to define a simple set of elements suitable for naive users to describe networked electronic resources. This was restricted to those needed for resource discovery of what were called *Document Like Objects* (DLOs). In the proceedings[Weibel95], the format of a set of 13 metadata elements, named the *Dublin Metadata Core Element Set* were defined by the participants:

> Subject, Title, Author, Publisher, OtherAgent, Date, ObjectType, Form, Identifier, Relation, Source, Language, Coverage

The elements are syntax-independent; no single encoding was defined and it was intended that they could be mapped into more complex systems such as SGML or USMARC[USMARC95] and can use any appropriate cataloging code such as AACR2, LCSH or Dewey Decimal.

Future Work

A prototype customisable interface to the archive is under development. Users can use a `FORM` and buttons in it, to describe how they want the presentation of the metadata, how rich, and what form. This would generate indices customised to the user and the browser, depending on what level of conformance it had to standards (unknown HTML, HTML 2, HTML 3, ...).

In addition to the collated indexers like ALIWEB and Harvest, there are web crawlers that try to *''index the web''*. These could benefit from rich metadata, provided by the document authors or site administrators that would be difficult to extract automatically. In the best of all worlds, each WWW site would create the metadata for each of the files it wants to make available to the world and the results would be distributed automatically using a hierarchy of caches (for efficiency). ALIWEB and Harvest allow forms of these kinds of systems to be built using IAFA templates and SOIF respectively.

Conclusions

A system has been designed for the Internet using IAFA templates as a basis. This has been very successful in organising a large archive (> 300 Mbytes with > 650 IAFA indices) of varied materials and providing good, detailed information about them. The indices for each of the access methods for the archive are created automatically and give a consistent look to the users.

The software can be found at the HENSA Unix archive at:

<URL:ftp://unix.hensa.ac.uk/pub/tools/www/iafatools/
<URL:http://www.hensa.ac.uk/tools/www/iafatools/

References

[Borenstein93]
N. Borenstein and N. Freed, *MIME (Multipurpose Internet Mail Extensions)*, September 1993, <URL:ftp://nic.merit.edu/documents/rfc/rfc1521.txt> and <URL:ftp://nic.merit.edu/documents/rfc/rfc1522.txt>.

[Boutell95]
T. Boutell and L. Wirzenius, *Linux Software Map*, June 1995, <URL:http://siva.cshl.org/lsm/lsm.html>.

[Bowman94]
C. Mic Bowman, P. B. Danzig, D. R. Hardy, U. Manber and M. F. Schwartz, *The Harvest Information Discovery and Access System*, Proceedings of the Second International World Wide Web, Conference, pp. 763-771, Chicago, Illinois, October 1994.

[Crocker92]
D. Crocker, *Standard for the format of ARPA Internet Mail Messages, RFC822*, University of Delaware, August 1992, <URL:ftp://nic.merit.edu/documents/rfc/rfc0822.txt>.

[Daniel95a]
R. Daniel Jr and M. Mealling, *URC Scenarios and Requirements*, Internet Draft, March 1995, <URL:ftp://nic.merit.edu/documents/internet-drafts/draft-ietf-uri-urc-req-01.txt>

[Daniel95b]
R. Daniel Jr and T. Allen, *An SGML-based URC Service*, Internet Draft, June 1995, <URL:ftp://nic.merit.edu/documents/internet-drafts/draft-ietf-uri-urc-sgml-00.txt>

[Deutsch95]
P. Deutsch, A. Emtage, M. Koster and M. Stumpf, *Publishing Information on the Internet with Anonymous FTP (IAFA Templates)*, IETF IAFA WG Internet Draft, January 1995, <URL:ftp://nic.merit.edu/documents/internet-drafts/draft-ietf-iiir-publishing-03.txt>.

[Hardy95a]
Darren R. Hardy and Michael F. Schwartz, *Customized Information Extraction as a Basis for Resource Discovery*, Technical Report CU-CS-707-94, Department of Computer Science, University of Colorado, Boulder, March 1994 (revised February 1995). To appear, ACM Transactions on Computer Systems.

[Hardy95b]
D. Hardy, M. Schwartz and D. Wessels, *Harvest User's Manual*, University of Colorado, Boulder, USA, April 1995, <URL:http://harvest.cs.colorado.edu/harvest/user-manual/>.

[Hoffman95]
P. E. Hoffman and R. Daniel Jr, *Trivial URC Syntax: urc0*, Internet Draft, May 1995, <URL:ftp://nic.merit.edu/documents/internet-drafts/draft-ietf-uri-urc-trivial-00.txt>

[Kopmanis94]
J. Kopmanis and L. Wirzenius, *Linux Software Map Entry Template*, August 1994, <URL:ftp://sunsite.unc.edu/pub/Linux/docs/lsm-template>.

[Koster94]
M. Koster, *ALIWEB*, Proceedings of First International WWW Conference, 25-27 May 1994, CERN, Geneva, Switzerland. ALIWEB is at <URL:http://web.nexor.co.uk/public/aliweb/aliweb.html>.

[McLoughlin95]
L. McLoughlin, `mirror`, Imperial College, University of London, UK, <URL:ftp://src.doc.ic.ac.uk/packages/mirror/>

[USMARC95]
USMARC Advisory Group, *Mapping the Dublin Core Metadata Elements to USMARC*, 1995, <URL:gopher://marvel.loc.gov/00/.listarch/usmarc/dp86.doc>.

[Weibel95]
Stuart Weibel, Jean Godby, Eric Miller, *OCLC/NCSA Metadata Workshop Report*, Dublin, Ohio, USA, March 1995 <URL:http://www.oclc.org:5046/conferences/metadata/dublin_core_report.html>.

Footnotes

[1]
HENSA Unix Parallel Computing and HPC Archive at <URL:http://www.hensa.ac.uk/parallel/>

[2]
HENSA Unix Archive at <URL:http://www.hensa.ac.uk/>

[2]
Linux archive, SunSITE USA at <URL:ftp://sunsite.unc.edu/pub/Linux/welcome.html>