2004

Why can't I manage academic papers like MP3s? The evolution and intent of metadata standards

James Howison
Carnegie Mellon University

Abby Goodrum

Why can't I manage academic papers like MP3s? The evolution and intent of Metadata standards

James Howison & Abby Goodrum
Syracuse Univeristy School of Information Studies
{jhowison,aagoodru}@syr.edu

Abstract

This paper considers the deceptively simple question: Why can't downloaded academic papers be managed in the simple and effective manner in which digital music files are managed? We make the case that the answer is different treatments of metadata. Two key differences are identified: Firstly, digital music metadata is standardized and moves with the content file, while academic metadata is not and does not. Secondly digital music metadata lookup services are collaborative and automate the movement from a digital file to the appropriate metadata, while academic metadata services do not.

To understand why these differences exist we examine the divergent evolution of metadata standards for digital music and academic papers. It is observed that the processes differ in interesting ways according to their intent. Specifically music metadata was developed primarily for personal file management, while the focus of academic metadata has been on information retrieval.

We argue that lessons from MP3 metadata can assist individual academics facing their growing personal document management challenges. Our focus therefore is not on metadata for the academic publishing industry or institutional resource sharing, it is limited to the personal libraries growing on our hard-drives. This bottom-up approach to document management combined with p2p distribution radically altered the music landscape. Might such an approach have a similar impact on academic publishing? This paper outlines plans for improving the personal management of academic papers—doing academic metadata and file management the MP3 way—and considers the likelihood of success.

1 Managing Music and Managing Electronic Papers

There is a striking difference in the ease and success with which people manage digital music compared to digital academic papers.

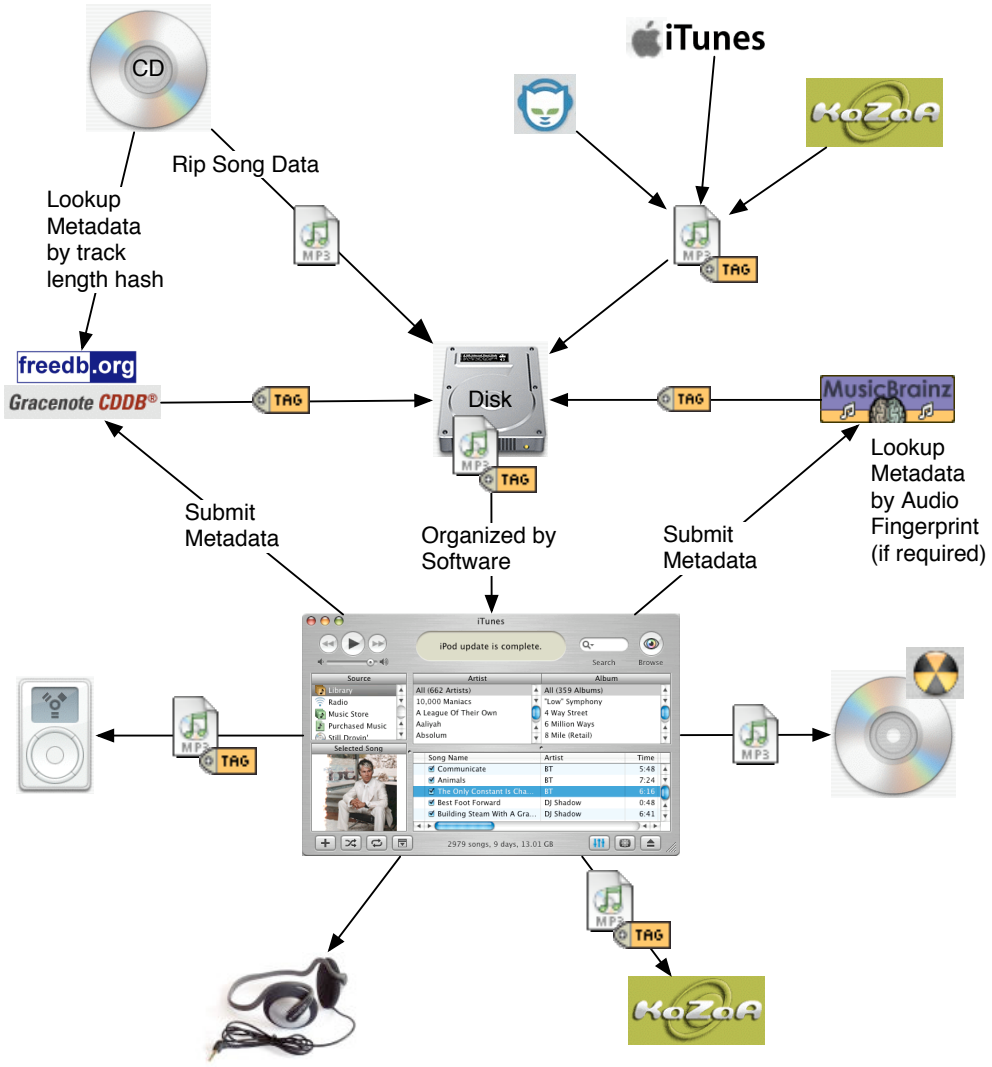

Figure 1: Workflow for personal management of digital audio

1.1 Managing digital music

Today's computers are filled with digital music and regular users are more than capable of meeting the exhortation to "Rip, Mix and Burn" [1]. They are able to bring music onto their computers, organize and file it for personal access, mix it into many combinations and burn it off to music CDs and digital music players. Even on the anemic interfaces afforded by today's digital music players users are proficient at managing substantial libraries of digital audio. Figure 1 shows the workflow for personal digital music management.

There are a range of programs available to assist the user throughout this cycle, many provided by the computer or operating system builders in an acknowledgment that managing personal libraries of audio is one of the factors driving the purchase of new computers. Programs like WinAmp and iTunes accept ripped CDs or files downloaded from online services, be they record industry

[1] Apple Computer's controversial marketing slogan emphasizing the ease of use of their digital music tools

endorsed or peer-to-peer sharing networks. In most cases the music is quickly and transparently stored on disk and presented to the user to be mixed and played by artist, album, genre or even beats per minute. Unlimited numbers of playlists provide personalized 'views' into the libraries. In addition to these 'kitchen-sink' music programs there is an entire ecology of standalone rip, mix and burners including the various peer to peer applications such as KaZaa and the venerable Napster (still kicking in the form of the OpenNapster networks). Through these networks, as well as LANs, burnt CDs and 'good old fashioned dragging from a friend's computer', these audio files are easily shared. In fact the controversy driven by this sharing reflects the sheer ease with which it occurs.

We venture to say that managing digital music provides probably the best personal information management experience available to individuals today, certainly exceeding managing email, word processing files, digital photographs and, central to this paper, personal collections of academic papers.

1.2 Managing academic papers

In fact the contrast between the management of digital music and academic papers is jarring. Firstly it is a much more uneven experience that is far from standardized and ranges from manual 'fumbling' through to quite sophisticated if users employ citation managers, such as Endnote or the variety of Bibtex management systems. Yet even the most sophisticated users experience difficulty in managing their files.

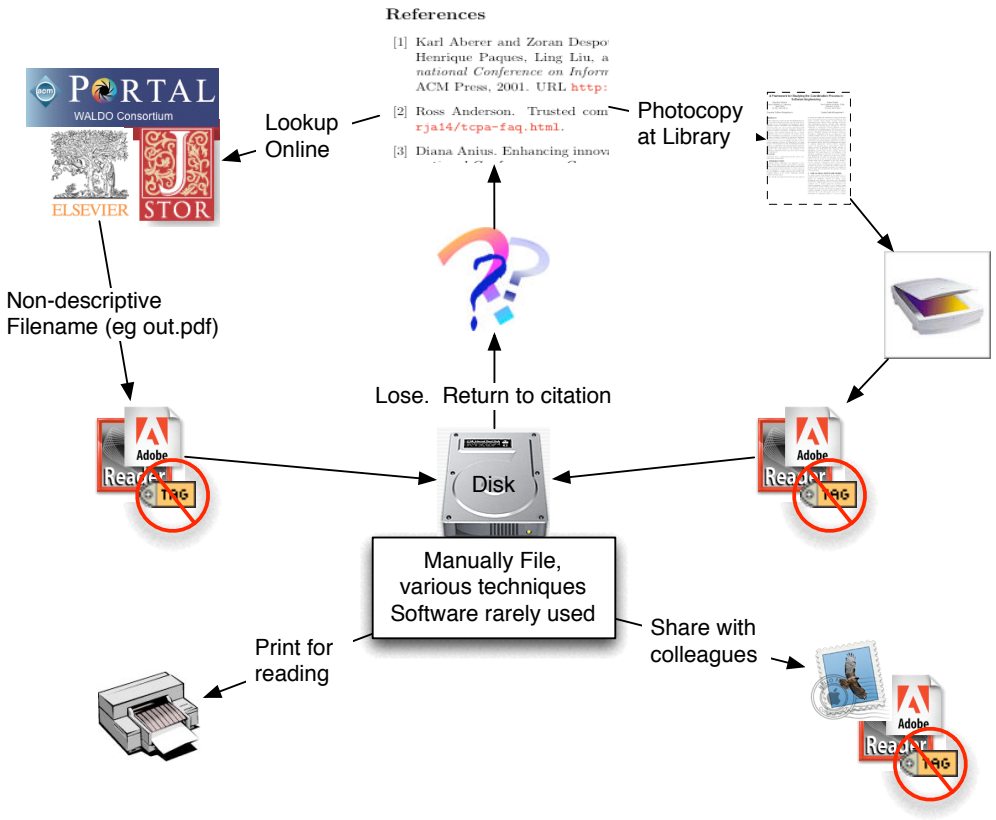

Figure 2: Workflow for personal management of Academic Papers

Academic workflow can be modeled, somewhat whimsically, as its own "Rip, Mix and Burn". Papers are 'ripped' from a variety of sources, mixed (and progressively 're-mixed') into annotated bibliographies, reading lists and eventually new papers which are 'burned' out, with their reference lists, to disks, paper and PDFs which are then circulated back to the academic community through email, websites and, hopefully for authors, conferences and journals. Figure 2 is an attempt to model the workflow of obtaining digital papers and Figure 3 attempts to model the use of digital papers in creating new academic works[2].

Our research into the personal digital management habits of academics is still in process: we are conducting interviews and a survey to document these practices. However for the purposes of this paper we offer scenarios which, while currently without empirical support, will be recognizable to most academics whether it reflect their habits or those of their colleagues.

The process regularly begins with a citation either garned from a colleague, a reading list or bibliography, a database or, very often, from the reference list of other papers. It is unlikely that the citation is in a structured database format, indeed quite often it is on paper or the digital paper of a PDF realized in the particular citation format of the publication.

Given just a citation most academics are quite adept at accessing academic papers from online journals, downloading, viewing and printing them. These printed papers might well then become the primary object, placed in piles by project and eventually filled in cabinets lining office walls—adding them to an existing system of photocopied papers obtained over a career.

Those academics that do choose to keep papers in digital form typically do so without the assistance of a document management system other than that provided by the hierarchical file system of the operating system itself. These might be filed by project or, more rarely we suspect, by author or personal genre taxonomy. In any case, so crude are the personal management techniques that it is often easier to store only the citation and return to the online repositories when the paper is again desired, despite that often being a technical violation of the terms of use [3].

The use of papers during 'mixing' (writing) might involve the use of a quotation which has to be typed from the paper copy into the document processing system being used, or copied (often with frustrating idiosyncrasies) from a PDF. Finally the reference list has to be prepared listing the sources used by the authors in their 're-mix'. Figure 3 is an attempt to model this workflow.

For those not using a citation manager this process can be quite frustrating and was described by one respondent as "fumbling".

Frustratingly, even having the original paper, in either digital or paper format, often is not sufficient to prepare the full citation. Rather earlier 're-mixes' (reference lists) are a common source of citations, manually transformed from the original published format into the intended outlet's required citation format.

Thankfully, there are a range of citation managers available in addition to graduate students!

[2]These workflow diagrams where published on the first Author's blog during the writing of this paper. Happily this prompted more sophisticated users to graph their own workflows and blog them. Links to these personal workflows which draw on a range of Citation Management software are available from the original blog posting via TrackBack. The post is at http://freelancepropaganda.com/archives/000302.html

[3]JSTOR For example states, "you may not download from the JSTOR archive an entire issue of a journal, significant portions of the entire run of a journal, a significant number of sequential articles, or *multiple copies of articles*." (JSTOR, 2002) (emphasis added).

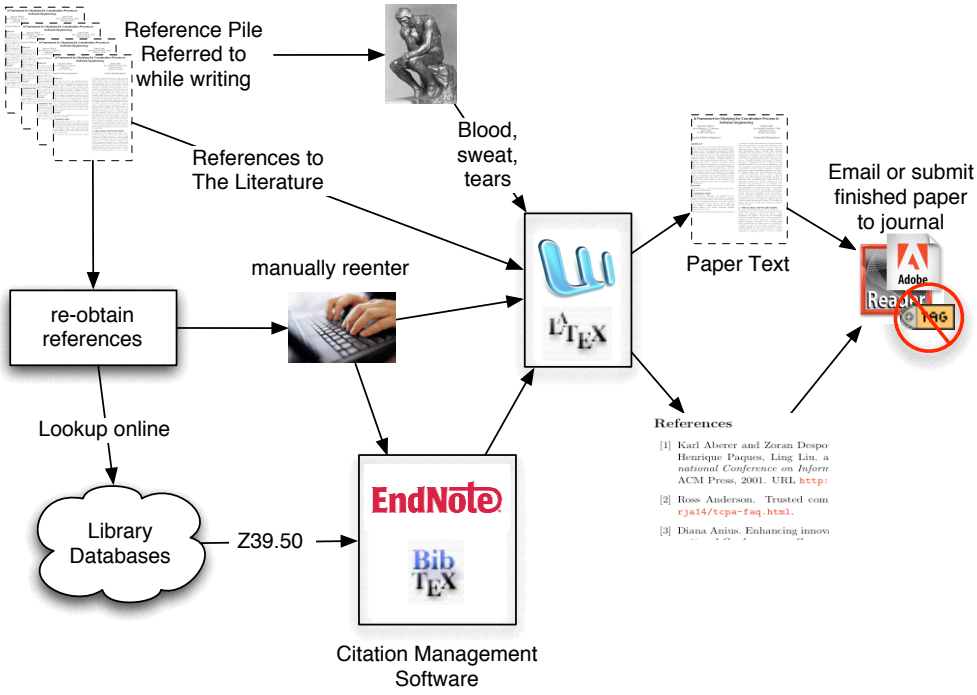

Figure 3: Workflow for writing Academic Papers

Programmatic citation managers, such as Endnote or Bibdesk, a Bibtex manager [4], integrate with the writing environment, such as Word or LaTeX, allowing the insertion of citation macros which are then used by the systems to prepare inline citations and reference lists in the specified format.

While convenient and powerful, these systems stop short of the positive experience of managing digital music.

The most frustrating thing about the academic workflow is that once the paper is shared, for example, by email or, hopefully, publishing in a journal, the entire process is repeated. The document access and citation effort expended by one author is almost completely unavailable to subsequent authors. The fact that my neighbor has recently found, photocopied and stored an article from our library is no help to my information needs, nor is their, often herculean, citation formatting. To be sure there are localized practices which facilitate shared repositories or citation exchange (Endnote and Bibtex files can be emailed or downloaded) but, it seems to these authors, these are not widespread, nor do they leverage the full potential revealed by collaborative peer-to-peer systems.

2 Why these processes are so different: metadata

Metadata is the key factor distinguishing the personal information management of music files from that of academic papers: the differences in the approach to metadata can significantly explain the differences discussed above. Firstly, music metadata is embedded within the file, while academic

[4]The first author is a sometime contributor the Bibdesk program which is an open source Bibtex manager. http://bibdesk.sf.net originally developed by Michael McCraken.

metadata is not. Secondly, metadata for music is obtained in such a way that later users can leverage the efforts of others while academic metadata is not. Metadata is important because it provides the raw materials used by tools to provide the easy digital object management and personal retrieval.

2.1 Storing music metadata

Music metadata is stored in the same file as the music data itself and is machine readable. MP3 files are the baseline standard for digital music and they utilize a standard known as ID3 which is inserted at the start or end of the file and can be read and written separately to the music data itself. More modern popular formats, such as AAC (MP4) and Windows Media similarly store their metadata in the file, using a file 'wrapper' able to handle many digital objects in the one file. MP3 ID3 tags are discussed in detail below.

With this metadata the management tools are able to provide many different 'views' of the personal library and are able to store and name the digital files in a logical manner under user control. There is no requirement for separate metadata databases.

The advantages of storing metadata with the digital object are twofold. Firstly embedding ensures that the metadata travels with the file content. Simply moving a file on a local disk does not separate or invalidate the metadata. More importantly moving the file across devices retains the metadata, thus downloaded music files often come complete with metadata. Similarly when files are moved to digital music players the metadata is available to facilitate collection browsing—without metadata these devices would be very hard to use.

The second advantage of embedded metadata is that the objects are never separated, if one is available the other is also available. They are thus far more readily available and less likely to get out of synch. This is particularly important for versioning e.g. when dealing with remixes or covers.

2.2 Storing academic metadata

By contrast academic citation metadata is managed as a separate object. This is true of 'paper' management as well as even the most sophisticated citation management tools available. There are a wide variety of formats for storing citation data but they are all used to create some type of database, from flat text or word files, to XML to SQL backends and proprietary binary formats.

At best these formats permit pointers to the actual digital object that they describe, for example providing a local file pointer, or a digital object identifier (DOI) or URL which points back to the publisher's canonical location for the object. They stop short of inserting the metadata directly into the PDF file.

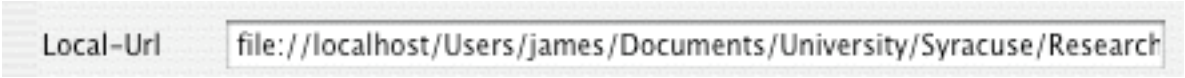

Figure 4: Bibdesk stores a reference to the file, but the file does not store the full metadata or even a reference to it.

2.3 Obtaining Metadata

Despite the advantages of embedded metadata for music there are two situations in which metadata is separated from the song content and must be added: when ripping from audio CDs or when downloading inadequately tagged files from p2p sharing networks. Happily there are automated techniques available for lookup in both cases.

The Audio CD format does not contain textual metadata. Users of standalone CD players will attest all that is available is the number of tracks, and their length, not the artist, album or and song title. Yet when a CD is inserted into a networked computer using any decent management software full metadata appears in the player, listing the song titles, artist name and album just as outlined on the CD liner. This metadata can be edited (to, for example, alter the genre) and can be retained through the ripping process and inserted in the digital tags.

This metadata is obtained through a network look-up which sends a hash of the file lengths to a server and receives machine readable metadata in return. The hash of an individual file length is not necessarily unique but that combined with the order of files on a CD (all provided in a CDs Table of Contents) provides a unique enough hash so that the CD can be identified with sufficient accuracy (freedb.org, 1999) [5] The network services that provide this metadata are the CDDB (now know as Gracenote) and FreeDB. They utilize a standard protocol, also known as the CDDB Protocol, to perform the hashes and file lookups (although GraceNote has transitioned to the non-standard CDDBv2). The evolution of these services is discussed in detail below.

Downloading music does not provide such a consistent experience. Most files on p2p networks have metadata embedded in the file, but a great many do not, or have inaccurate or useless metadata (it is not uncommon to see the name of the person that ripped or shared the file in the metadata tag e.g. "ripped by l33tboy"). So haphazard is the provision of metadata that this has become a point of commercial differentiation for the legal download services claim accuracy and completeness and typically include at least the album cover which good music players can display.

Even when a music file is downloaded or ripped from a CD without metadata, there are network look-up services that can automate the its discovery. These are relatively new and utilize the actual musical qualities of the songs to generate a hash for lookup, rather than the track length and order. They are therefore very flexible when dealing with the inevitable differences between tracks ripped by different encoders or with differing lengths, although the trade-off is that there are many more near collisions and thus more user selection interaction required. The best known is probably MusicBrainz. MusicBrainz can be run over a collection of digital music files and, provided the track is in the database, it will return a number of suggestions for the track and add the appropriate metadata to the music file. The tagger runs over a digital music collection and conducts an exchange of metadata adding metadata from the server to tracks without it and uploading metadata from tracks on the local computer that are not yet present on the server. MusicBrainz does not store the actual recordings. Figure 5 compares these two music metadata processes and emphasizes the collaborative their collaborative nature.

[5] Although collisions are theoretically possible the authors have never witnessed one. In any case if more than one album matches a simple interactive user selection can resolve the conflict.

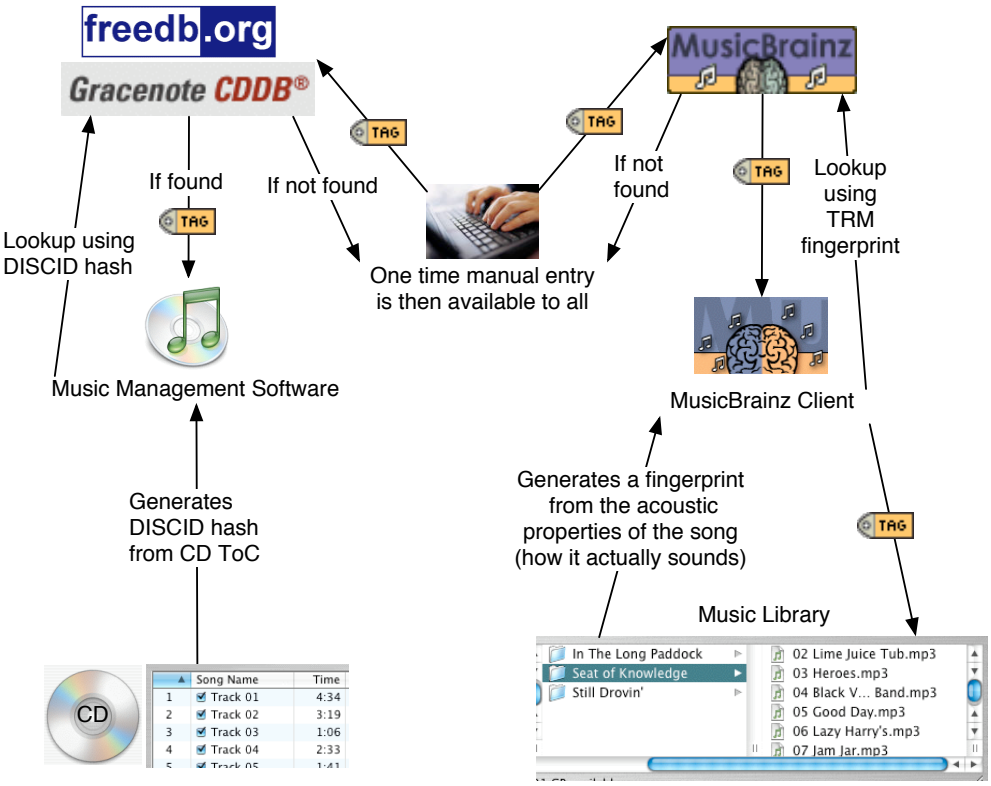

Figure 5: Accessing music metadata: both ways of accessing metadata for digital music leverage a single submission to provide many lookups—they are collaborative services

2.4 Obtaining Academic metadata

Academic metadata is obtained in a variety of ways, yet even the most sophisticated methods stop short of the level of automation and convenience reached by digital music. For those not using citation management software the process is a manual one of re-entering citations in the required format. This is often done through "reference list raiding", is subject to copying errors and is extremely tedious.

Thankfully there are ways to access machine readable citation data other than manual entry. Citation managers can connect to, search and import metadata from bibliographic databases that support the Z39.50 protocol. Efforts are underway to provide Z39.50 access through a standardized programming interface, known as ZOOM. Vendors, such as OCLC, Ovid, and NISC, offer a direct export feature which allows you to automatically transfer your saved citations into your citation management software. The metadata available for searching is sparse, however, and in most instances, the connection must be handled through an institutional account or proxy server. For those databases that are not Z39.50 compliant or for complex searching, one searches using the native database search functions, then downloads the data to a text file and imports that file into citation management software using an appropriate filter configured for each database/vendor if one is available.

Some online resources providing access to academic papers do carry the citation metadata with

their referent papers either as text or as exportable files from the vendor's bibliographic database. Many, however, do not provide any formatted or machine readable metadata along with their PDF images—in fact most online services do not even use a usefully descriptive file name, often defaulting to out.pdf" or what is presumably an internal database number but of no use to the user like "4456.pdf". The full text of a paper is downloaded as a PDF from one resource and the citation metadata is downloaded into citation management software from another resource. The changing nature of scholarly publishing in some disciplines has also led to a more informal, non-archival, publishing model by individual and institutional authors in HTML, PDF, Latex, or Postscript formats on the web. In most of these cases, citations and other important metadata are not carried with the papers or even provided from a linked bibliographic repository.

These more sophisticated systems still involve substantial manual work on behalf of the researcher. The 'keys' like title and author, used for lookup are part of the citation themselves and there is no widely used automatic way to move from the digital object, a downloaded PDF, to the metadata. That is, there is little that is analogous to the DISCID hash or the acoustic hashes used by MusicBrainz. Instead search is limited to a search key, like title, extracted by individuals—and this process is repeated by each individual that wishes to access the metadata. While machine addressable repositories of metadata do exist the lookup process is not automated nor does it leverage the work of others in linking a paper with its metadata or allow end-users to contribute missing metadata.

One partial exception is Autonomous Citation Indexing (Lawrence et al., 1999*a*). This process uses AI techniques to probe the content of a digital file for its own citation metadata, scanning for elements like authors and titles, drawing on similarities in document formatting. This system has been used to build the 'grey-literature' computer science database, CiteSeer, which is discussed in more detail below. Autonomous Citation Indexing by itself does not, however, leverage the neccssary manual corrections and additions from outside the document made by other academics in the course of their writing work—although, as discussed below, CiteSeer has built a partial solution.

The academic metadata lookup process, therefore, lacks the strong connection between content and metadata that facilitates the effectiveness of the digital music process. Because metadata is managed as a seperate digital object obtaining it is an interactive process which requires seperate user involvement and searching skills. This seperation is also a source of uncertainty about the version correlation that is the bane of electronically received academic works.

The availability of metadata is the factor that facilitates successful management digital resources. It is easier to accomplish obtaining and managing of metadata for digital music because of the combination of storing metadata with the file and the availability of collaborative services that support retrieval of metadata given the digital object.

By contrast academic metadata is stored outside the file and the movement from paper or file to machine readable metadata is not supported by an automated system. Therefore citation management is largely a difficult and repetitive process that fails to leverage the efforts of others.

There are, however, the beginnings of a better way of doing things and this paper attempts to draw the threads together below. First, however, we outline our main contention: an explanation of the reasons for these differences can be found in a comparison of the intent and evolution of these two differing metadata schemes.

3 How did it come to this? The intent and evolution of metadata standards

Music metadata was designed for personal file management and with the expectation that users would first have the object (CD, file) and then seek the metadata. Academic metadata was designed for information retrieval and archiving and with the expectation that users would first have the metadata and use that to access the object. These differing intents largely explain the situation we face today.

3.1 Evolution of the ID3 tag

The ID3 metadata tag was born as a standard for personal information management and only later became used for information retrieval.

The development of the standard was motivated by the frustration of users seeking to manage the MP3 files they were creating. The MP3 data format, as designed by the Frauenhofer institute originally had a limited scheme that only facilitated the storing of information that was publisher relevant: e.g. copyright, copy/master bit and a 1 bit field called 'private' which was for application specific use. These facilities, while seen as useful from the publisher's perspective, did not facilitate personal information management. Indeed the design of the MP3 header format was influenced by the requirement that it be suitable for broadcasting, rather than local play or management (Bluecygnet Technologies, 2001; Alpha Internet Consulting, 2000).

In 1996 this difficulty lead Eric Kemp (alias NamkraD) to develop simultaneously the ID3v1 standard and a tagging programme that read and wrote the tags called "Studio3" Nilsson (2000). The program was for personal file management. The ID3v1 standard provided the basics of metadata and facilitated the organization and 'mixing' of music files. We interviewd Martin Nilsson, who was to develop the second version of ID3, who stated that when he first encountered MP3 tagging "was [a way] of keeping things neat and 'sophisticated". He added that tagging wasn't used in early MP3 retreival systems, "I do know that MP3Get [a campus searching program used by the ID3v2 developers] didn't have any metadata, and it never really mattered in practice." (Nilsson, 2004).

The ID3v1 standard was limited in ways that frustrated users seeking to manage their personal music libraries. For example the length of information that could be stored in each field was limited to 30 bytes and plain ASCII and there was no ability to add additional fields. An important frustration was that the tag was at the end of the song, reducing its usefulness in situations where the track was being streamed (because the tag would arrive last). These difficulties lead Martin Nilsson to design a new version of the tag which is known as ID3v2 and is still in use today (Nilsson, 2000). ID3v2 was rapidly adopted by the booming digital music industry (including developers of software and companies trying to become publishers of digital music). The first formal release of the tagging standard was at the MP3 conference organised by Michael Robertson's mp3.com (Nilsson, 2004). The ID3v2 developer's attendence at the conference was sponsored by mp3.com and e-music.com

To this day the ID3v2 standard is only an "informal" standard—it has not been adopted as an official standard of any standards organisation, although experience from its development has been applied to the MPEG-7 efforts. Nilsson argues that it was not required as the adoption was sufficient

and the process time-consuming, "I did a 'soft start' with the MP3 content type RFC (now RFC 3003), and it took such an enormous effort to get that document in place that I thought that it really wouldn't be worth the effort to get ID3v2 into an RFC. It was already well known." (Nilsson, 2004). A now expired Internet Draft (precusor to an RFC), draft-nilsson-id3-oo.txt, was written in December 2000 and the official spec is hosted on http://www.id3.org. This is in striking contrast to the heavy standardization effort familiar to those engaged with academic metadata.

There are three striking elements in the evolution of the metadata standard for digital music. Firstly the standards were developed by end-users seeking to facilitate personal uses of the music files they were creating. Secondly the standard was created at the same time as its implementation and was designed to meet just those goals required at the time. Thirdly, the standard was propagated freely and without serious efforts to achieve an official status for the standard.

Compared to the evolution of metadata standards for academic papers, outlined below, the ID3 tag had a simple intent, had significant 'first mover advantage'—there was no competing standard—and was in a field that lacked an obvious standards forum. The open culture of sharing which exists in the online music community promoted interoperability and there were no incentives to 're-invent the wheel'.

3.2 The evolution of collaborative metadata services

Inserting a CD into a computer in the mid-90s wasn't very different from using a CD player. In particular it was likely that only the track numbers and lengths were available to the user and that track selection occurred by manually cross-referencing the physical liner notes. The situation is different today because of collaborative work by internet users that resulted in a service called CDDB (the Internet Compact Disc DataBase).

In a story that recalls the history of the Domain Name System, the CDDB has its origins in a text file that resided locally on a user's computer in which metadata that had been manually entered by users was stored. Also stored was a hash that allowed the player to remember if a CD had been seen before and to associate the appropriate metadata. By 1996 the internet had facilitated trading and combination of these individually entered files, so that the file grew so much that it began to exceed the ability of Windows players to use it MusicBrainz (2004). The solution was the development of a client/server database hosted at cddb.com. The protocol developed for this database allowed both the retrieval and, crucially, the submission of CD metadata by individuals.

The database grew quickly through these individual submissions providing a valuable service to the digital music community. It was peer-to-peer in that the service facilitated the transfer of digital information entered by one user to many others. However the CDDB actually stored copies of the object being shared, the metadata, on a centralized server—by contrast Naspter, which also utilized a centralized server, did not store the actual object, just a reference to which user's hard-drive it could be obtained from. Like other centralized collaborative systems, however, the CDDB was vulnerable to attack. Yet unlike the legal attack on Napster this attack came from 'inside' when the CDDB was commercialized by the owners of the server who took out a patent on the system and announced their intention to begin to charge the authors of software (as opposed to the users) for their software's access to the database.

Unsurprisingly this action caused a uproar on the behalf of the users who had voluntarily contributed the contents of the database and who had expected reciprocation in exchange. FreeDB is

the open source alternative developed as an alternative to the the CDDB service.

However CDDB had a strong head start and their strategy of charging small fees to the authors of music management software has proven to be the foundations of a reasonably successful business. They renamed themselves to Gracenote and developed a patented new protocol, CDDBv2 which effectively locked free players out of the community contributed database. Gracenote did, however, face the prospect of weeding out 'junk' or misspelled data from their system.

The focus of CDDB reflected the state of digital music in 1996. It is a system that is tightly coupled with CDs themselves and was initially developed without an expectation that the songs would necessarily be 'ripped' from the CDs. The system did not use require a tagging format—the metadata was stored separately from the file. Tagging was left to the MP3 upstarts lead by Eric Kemp and eventually Martin Nilsson. In fact the CDDB was not interested in an engagement with the mp3 tagging efforts: Nilsson paraphrases their response as "We don't want to be associated with MP3, since record companies think of it as illegal".

The rapid growth of digitized music and the p2p file-sharing applications outpaced the ability of music users in maintaining accurate metadata. Enormous downloaded collections were without useful metadata for personal management other than filenames, or have polluted metadata. Recently other services have emerged to make the link between music files and the metadata needed to manage them in large collections and on digital music players, like the iPod. These services cannot rely on the firm publisher determined CD publishing format of track length and order, because files are individualized and ripped in ways that alter their length.

The alternative strategy now adopted is more sophisticated. It relies on acoustic fingerprinting algorithms that create a hash of the way the music actually sounds. This also opens the service up to music never released on CDs such as audience live recordings and underground 'mash-up' remixes. MusicBrainz also solves the 'dirty' data problem through collaborative voting which reflects the community's opinion on which metadata is the most accurate and/or complete.

Digital music metadata is widely available today because collaborative services, p2p in function if not in technical structure, allow large groups to leverage the minimal efforts of individuals in manually entering the music metadata. These services combine with the development of embedded tags that store the metadata within the file to facilitate the effective digital music experience widely enjoyed today.

3.3 Evolution of Academic metadata standards

Academic papers and metadata have historically been managed as separate objects. A separation of object and metadata made sense before the development of computer systems capable to benefit from their combination— placing the catalogue cards with the book on the library shelf would not have been a sensible system! Today however the continued separation of metadata and object is the cause of significant frustration and inhibits building management applications.

Metadata in the form of indexes are as old as publishing itself, coming into their own as stand-alone entities with the rise of scholarly publishing in the 18th century. As publishing began to computerize their processes, abstracting services began to provide access to their publications' metadata-only digital catalogues on CD-Roms and eventually migrated to dial-up and web-based access. It is important to note that in most cases, access to the bibliographic record or to an abstract did not also entail access to the full-text of scholarly papers. Full-text availability is still

a new phenomena in academic publishing and many journals are still not available in full text outside of print. Moreover, the publishers of bibliographic indexes, abstracts, and catalogues have not been the publishers of the journals themselves. For these reasons, academic metadata has long been provided apart from the papers themselves. The same historical circumstances explain why metadata lookup systems are not collaborative nor complete—they have been used for comparative economic advantage.

Today however this continued separation of metadata and object is the cause of significant frustration and inhibits building effective management applications and sharing the work of linking metadata with the digital files.

It is clear that the metadata required for personal file management differs between music and academic resources. Academic metadata is more complex than music metadata. While artist, album, title and genre is sufficient to describe the vast majority of music academic metadata schemes must deal with quite a number of different publication types requiring different schemas, eg Journal Articles, articles in conference proceedings, books and book chapters to name but a few. No doubt this complexity is a factor in the lack of standardization of academic metadata formats. A recent review of academic metadata for personal bibliographic management argues, however, that "While the library standards are largely 'overkill' with respect to personal and small group management of data to produce a bibliography, many of the basic needs behind the requirements remain the same." (Shabajee, 2003). When academic metadata is considered only as those fields required by academics when preparing bibliographies the complexity can be significantly reduced. Schemas like `bibtex` and `refer` (used by `Endnote`) adequately serve the needs of large communities. Finally, additional complexity has been introduced to academic metadata systems by their information retrieval focus, information that is not used in the paper writing format.

4 Doing academic PDFs the MP3 way: A better way?

How then can the lessons of digital music management be applied to improve the practices of personal information management by academics? The two key differences identifed in this paper are: storing standardized metadata with the file and collaborative metadata lookups. The proposals developed here are modest and minimal and aim at bottom up adoption by emphasising immediate usefulness for individuals managing their collections of digital academic papers.

4.1 Getting metadata into PDF files

The Portable Document File (PDF) is an open standard defined by Adobe and implemented in their Acrobat product line. The openness and fidelity to paper of this standard has lead it to become the ubitquious format for electronic publishing of academic papers. Free readers have been available for upwards of 8 years. Adobe charges for the ability to create PDFs providing sophisticated Macros for Microsoft and their own page layout products. However there are an increasingly number of free and open source products also able to produce publication ready PDF files, for example the LaTeX system which initially produced PostScript now typically writes straight to PDF and the Apple OS X operating system uses PDF as its screen-display description language and is able to produce a PDF from any document or screen-shot.

The PDF standard has the ability to store a pre-defined and limited set of metadata in the "Docu-

ment Information Dictionary" (DID). The DID has fields only for title, author, subject, keywords, Created on, Modified On and Producing application. Perhaps because of these limitations, or perhaps because of limitations in production tools, these fields are extremely rarely used in the academc world. None of the academic electronic document publishers used by the authors utilizes these inadequate fields.

Adobe has produced a relatively new open standard for storing metadata which can be used for PDFs and in a wide variety of file formats. The eXtensible Metadata Project (XMP) allows the addition of "bullets" of well-defined data in the XML-like Resource Description Format (RDF) Adobe Systems Inc. (2004). This has excellent potential for reaching ubiquity as the cross-platform container for machine readable metadata. Adobe has provided a Software Development Kit but currently it is very difficult to produce compliant XMP and add it to PDFs without using the for-pay full Adobe Acrobat Libraries. This is slowing adoption but the open source community is committed to building tools to utilize XMP—it is a highlighted 'Tech Challenge' of the Creative Commons project Creative Commons (2003).

4.2 Is it necessary to standardize machine readable metadata?

If XMP is the candidate most likely for bringing together the digital academic paper with its machine readable metadata, what format is that metadata to be in? The metadata required for personal and citation management is orders less complicated than that required for library use but still more complicated that for music. Countless hours have already been spent in committees aiming to standardized academic metadata.

Happily it is not necessary to adopt or promulgate only one standard—indeed the recent heated debates over the format for blog feeds (RSS versions and Atom) indicates that many variations can co-exist Ecker (2003). Because the formats are both written and read by machines and because RDF formats reference namespaces available online, variations in formats are not the great problem they once seemed. The next-generation Z39.50 projects are quite liberal in the formats that they can deal with, including `marcxml`, `mods` and `dc`. In fact, while this paper has highlighted the ID3v2 tagging format, there are quite a number of others hidden behind the scenes, including OGG-tags and ACC-tags. In the first instance it seems reasonable to employ RDF translations of the widely used Bibtex and Refer formats but the authors are actively researching this topic.

4.3 Making academic metadata a collaborative practice

Our discussion of digital music emphasized the role that collaborative metadata databases, such as CDDB and MusicBrainz, play in maintaining the ease of the digital music process. We also argued that two elements needed to be resolved before the excellent start that is Z39.50 can be as useful in the management of academic digital papers. The first was a way to automate the lookup process so to easily move from the digital papers to the machine readable metadata. The second is to leverage the efforts of one's peers to build comprehensive databases.

The more sophisticated approach of MusicBrainz and other acoustic matching services points the way forward. Recall that these services use a hash built from the acoustic properties of the digital music. Analogously, a lookup hash can be harvested from the actual content of the paper. In fact there are many algorithms available to measure the textual similarity between documents, all that is needed is to use these to match downloaded PDF files with metadata in a digital repository.

These 'fuzzy-matching' techniques have two added advantages over the more simplistic approach of calculating a simple hash from the PDF file (e.g. MD5). File hashes are excellent at assessing whether files are identical but even semantically meaningless changes render vastly different hash results.

Content based hashes do not have that limitation and can, in fact, help resolve the versioning problem. If a user has a digital copy of an earlier version of a paper the content matching systems should be able to match that with its more developed 'kin' and return the results in publication order, thereby alerting the user to a later version that might be of great interest.

One challenge to the content based lookup approach is that many journal services provide their PDF files as images, rather than text. For older articles that have, in fact, been scanned from paper copies the reason is obvious. For newer articles which are generated from the digital production files the reason is, to these authors, obscure. The approach that seems most promising is the application of optical character recognition techniques to the image documents. It is an open research question to find content analysis algorithms that are able to make accurate matches even in the face of the noise introduced by faulty OCR.

The second requirement is to inject collaboration into the process of citation management by allowing users to upload the metadata that they associate with a paper and to make that available for query by others. The existing metadata databases are publisher and institution bound and are therefore incomplete—worse, they are viewed as competitive assets and obstacles are placed in the way of their consolidation. Peer-to-peer collaborative uploading appears to be the most likely way to build an adequately large collection of downloadable metadata.

We are actively researching the creation of such an automated collaborative metadata service and welcome any input regarding the components outlined above.

4.4 Prospects of success

Incentive issues must always be considered when building collaborative systems. The system would not be of great use if no-one was motivated to associate metadata with a digital file and upload the hash and metadata for others to download. It is arguable that the success of digital music sharing is the result of the high and widespread demand for the product—and the time available to the often teenaged participants. These conditions don't hold for academic papers.

Thankfully systems like the CDDB or the one proposed above have the characteristic that only one upload is required to make the metadata available to all. It is established that online availability increases citations to one's work so it does not seem a stretch that easy availability of accurate metadata would also increase citations Lawrence (2001). Therefore the candidate with the greatest motivation to provide metadata will be the authors themselves. In any case the effort of preparing metadata is performed by every academic who chooses to cite a work and the marginal effort required to link this with the file and upload it is quite minimal—and likely to be of personal use to the up-loader next time they seek to use the file. For these reasons we do not consider there will be a lack of motivation to upload metadata to the system.

Checking this metadata for accuracy is an important task and it is again one that we expect in the medium term to be primarily provided by the paper's authors. The Citeseer service provides Bibtex metadata which is in the first instance extracted using automous citation indexing methods but which is also open to community modification (Lawrence et al., 1999*a*) . Anyone who has a free

Citeseer user account can modify and edit the metadata available through the site (Lawrence et al., 1999*b*). This openness there has lead to a clear trend of increasing accuracy and completeness of metadata on the Citeseer service, proving resiliant, thus far, to pollution.

5 Conclusion

The personal information management of academic papers has a long way to go to approach the ease and usefulness achieved in the management of digital music. Yet the processes are not that distinct. We have argued that two crucial differences that hold back personal information management for academics are that the metadata and content are managed as separate objects and that the metadata lookup services are not collaborative.

Management practices for digital music evolved from the need to manage growing libraries of 'ripped' music, only later did metadata begin to play a role in information retrieval. In contrast, academic metadata was designed primarily for Information Retrieval and there has been very little focus on its use in personal information management.

We propose that two tasks are required to move towards more effective practices: using XMP to store academic metadata with the file and using content-based hashing to map between the digital files and a user contributed metadata database.

Our next step will be to conduct an empirical investigation of the way academics obtain and manage their digital papers and citation preparation processes. We also intend to prepare a proof of concept for a collaborative metadata database that can be searched using content-based hashes.

References

Adobe Systems Inc. (2004), 'The extensible metadata platform'. Available from: http://www.adobe.com/products/xmp/.

Alpha Internet Consulting (2000), 'Inside the MP3 codec - MP3 anatomy'. Available from: http://www.mp3-converter.com/mp3codec/mp3_anatomy.htm.

Bluecygnet Technologies (2001), 'MP3 Frequently Asked Questions'. Available from: http://www.bluecygnet.com/amce/faq-amce.htm.

Creative Commons (2003), 'Tech challenge: Open source application support for xmp'. Available from: http://creativecommons.org/technology/challenges#challenge_entry_4005.

Ecker, C. (2003), 'The great syndication wars (RSS and Atom)'. Available from: http://phaedo.cx/archives/2003/07/15/the_great_syndication_wars.html.

freedb.org (1999), 'Discid howto (generating cddb hashes)'. Available from: http://www.freedb.org/modules.php?name=Sections&sop=viewarticle&artid=6.

JSTOR (2002), 'JSTOR Terms & Conditions of Use'. Available from: http://www.jstor.org/about/terms.html.

Lawrence, S. (2001), 'Online or invisible?', *Nature* **411**(6837), 521.

Lawrence, S., Bollacker, K. & Giles, C. L. (1999*a*), Autonomous citation matching, *in* O. Etzioni, ed., 'Proceedings of the Third International Conference on Autonomous Agents', ACM Press, New York.

Lawrence, S., Giles, C. L. & Bollacker, K. (1999*b*), 'Digital libraries and autonomous citation indexing', *IEEE Computer* **32**(6), 67–71.

MusicBrainz (2004), 'MusicBrainz roots and history'. Available from: http://www.musicbrainz.org/history.html.

Nilsson, M. (2000), 'The short history of tagging'. Available from: http://www.id3.org/history.html.

Nilsson, M. (2004), 'Email interview with id3v2 developer'.

Shabajee, P. (2003), 'Review of personal bibliographic systems'. Available from: http://www.ilrt.bris.ac.uk/publications/researchreport/rr1032/report_html?ilrtyear=00.