



> Semantic Web Use Cases and Case Studies

Case Study: Establishing an Open, Digital Media Commerce Standard Using Semantic Web Technologies

Manu Sporny, Dave Longley, Mike Johnson, and David I. Lehn of Digital Bazaar, Inc., Blacksburg, Virginia, USA

December 2008

Introduction

This document outlines how Digital Bazaar is using Semantic Web Technology to establish a set of open mark-up and communication standards for Web-based, peer-to-peer marketplaces. The system that Digital Bazaar has created, called Bitmunk, is used to transact digital media such as music, movies, television and books between independent agents on the Web. The decentralized nature of the peer-to-peer marketplace requires flexible, open standards for communication and knowledge representation.

The Digital Media Commerce Problem

At present, digital media sales on the Web is largely a producer to consumer model. Typically, sites like Apple iTunes, Amazon Music, Netflix, the Microsoft Zune store and other digital retailers create Web-based store fronts to sell digital products to their customers. During the product sale, the item is either wrapped in Digital Rights Management technologies in an attempt to curb piracy or licensed only to the individual purchasing the digital good.

There currently is no secondary market for digital content sales due to the nature of digital media - reproducing a digital file is incredibly fast, cheap and easy. This fact has led to a great amount of digital piracy, notably by peer-to-peer network technology such as Napster, BearShare, Morpheus, Gnutella, Kazaa and BitTorrent. While iTunes has claimed to make as many as 435,164 video sales per day through their store, the number of video downloads per day over Mininova, one of the largest BitTorrent networks, approach close to 4,975,075 downloads per day. Close to ten times as much video traffic is served through a single BitTorrent community versus the number two digital video sales service in the world.

While it is still fairly unknown how much DRM, RIAA tactics, value proposition, platform access, file formats, convenience and digital kleptomania contribute to piracy - it can be said that each contribute to piracy in today's digital media ecosystem. One of the goals of Bitmunk is to address the following key areas that are currently overlooked by the major players: value proposition, platform access, file formats, and convenience. The end-goal is to create a peer-to-peer network, driven by open, flexible standards, that allow customers to help artists, production companies and creators of content to distribute digital goods in exchange for financial compensation. Just as physical storefronts help to distribute products on behalf of their manufacturers, digital storefronts can help to

distribute content on behalf of their creators.

Some of the problems we faced when creating an open standard for buying and selling digital content is not only ensuring that the system is stable and capable enough to perform millions of digital/financial transactions per day, but also flexible enough to expand as different digital content types and business models are created. There were a number of design questions that we faced:

- How does one create an easy-to-author media markup format that integrates well with the Web?
- How does one create a digital commerce markup format that everyday people could publish in their blogs?
- How does one create an enforceable digital contract embedded with strict legal meaning?
- How can we ensure that the markup mechanism is in-line with the future direction of the Web?
- How would the peer-to-peer network interact with the Web such that the end-user experience is minimally invasive, if not automatic?
- How could Web-based agents interact with the data to provide a richer browsing experience?

The Long-Term Goal for Bitmunk

Bitmunk has been created to provide a legal mechanism to buy and sell digital content via peer-to-peer networks. The system is designed to operate on a variety of devices including mobile phones, pocket PCs, netbooks, laptops, set-top boxes and desktop computers. The end-goal is to provide an open, extensible, standards-based network for digital content sales.

Creating a Web-based Music Representation Format

A Brief Introduction to Representing Music on the Web

The gold standard for identifying musical works stored by computers has been the [ID3](#) tagging format used in [MP3](#) audio files. These files have been traded, bought and sold across the Web for years without a major change to the underlying representation format. As our data transitions from the desktop onto the web, new methods of interacting with that data are being created. This section focuses on how we use the music data that we have on our computers and how we can express that data on the World Wide Web.

[Digital Bazaar](#) operates a service called [Bitmunk](#). The purpose of the service is to connect artists with fans and then help the fans distribute the artist's digital creations. More specifically, the service provides an [MP3 song catalog](#) of over 50,000 independent musicians and over 850,000 songs. When a fan buys an artist's album, they can re-distribute it on the artist's behalf via the Web. This allows the fan to legally re-sell the artist's album through the fan's website or blog. The artist is paid a royalty and the fan is paid a commission for helping to distribute the work on behalf of the artist. This is a fundamentally different approach to music distribution, focusing on rewarding good behavior rather than punishing bad behavior.

One of the biggest hurdles facing the company is helping fans find independent musicians that they enjoy. Recommending new bands is a difficult task because the music selection process is a chaotic one. Music tastes almost never form immediately and are not greatly influenced by any single website. Friends, advertising, radio stations, blog loyalty, and many other factors play into

what a fan likes and where they go to get their music.

Bitmunk needed a mechanism where someone could recommend music via their blog. The recommendation would be marked up in such a way as to be universally read-able by any Web browser. The "music object" could be used to pull more information from the Web without requiring the person browsing to go to another page. Information such as artist bio, related album information, related music blogs, purchase information and band history could be pulled in the background while browsing a page. More importantly, the means to enable bloggers to not only recommend music, but sell music directly from their blog was a strong requirement for Bitmunk.

The representation mechanism had to have the following properties:

- A way of representing music and other digital media on a web page so that web browsers could recognize albums, videos, actors, artists and other media metadata.
- A method of creating community standards for representing music, movies, television and books.
- A method of representing commercial transactions including prices, and currencies that were attached to digital goods.
- Broad web-browser integration.
- Ease of publishing, tool creation and a potential for rapid adoption.

What follows is the story of what our company tried over a 20 month period, the mistakes we made, the successes that we have enjoyed and hopefully, some guidance for others trying to do the same thing in their industry.

The Promise of the Semantic Web

There have been many different postulations about what the [Semantic Web](#) is and what it can accomplish for human kind in the next 20 years. Some believe that there will be a Semantic Web that runs parallel to the [World Wide Web](#). Where the World Wide Web is for people to read and understand, the Semantic Web will be for computers to read and understand. Both Webs will talk about the same thing, but in a way that their respective readers can understand. Much like most British citizens read newspapers written in English and most Brazilians read newspapers written in Portuguese - the Semantic Web may be partitioned such that only computers can understand it. Others believe that the current World Wide Web will morph into the Semantic Web by inserting machine-readable bits and pieces into the web pages that we are accustomed to today.

While there are many theories and scenarios that academics have postulated, Digital Bazaar focused on a number of very practical questions:

1. "How can we teach a computer how to recognize a song and an artist on a web page?"
2. "How can we express that a piece of digital media is available for sale from someone's web server?"
3. "How do we do it in a way that novice web publishers can understand?"

These questions were at the heart of the work that Digital Bazaar has been performing for the past 20 months. These questions guided us through the possibilities and helped us focus on the practical.

Technology Requirements

When the company started to look at technologies and projects that could help Digital Bazaar achieve its goals, several general requirements were outlined:

- **Simplicity** - The end solution had to be simple to understand, implement and author. This would help rapid adoption of the technology.
- **Standardization** - The mark-up of music metadata had to be a standardized format because the company wanted wide adoption of the standard. Digital Bazaar believes very strongly in standards being The Right Thing To Do when creating any large electronic ecosystem.
- **Distributed Innovation** - The implementation mechanism had to be distributed in nature to ensure innovation and survival of the standard outside of Digital Bazaar. Our employees want to be able to continue innovating, even if Digital Bazaar ceased to exist.
- **Re-use** - The company would re-use grammars, data, technology and concepts that had widespread use. Our employees wanted to be focused on the end-result and not theoretical debates on what the future might hold. There was a specific problem that needed to be solved and the company would choose to focus on that and not the latest technology bandwagon.

With those general concepts in mind the following technologies and projects were identified as interesting:

- **Public Music Metadata** - [MusicBrainz](#), [Discogs](#) and [Freebase](#)
- **Private Music Metadata** - [CD Baby's](#) artist catalog.
- **Standardization Communities** - The [IETF](#), The [World Wide Web Consortium](#), [OASIS](#), and [Microformats](#).
- **Previous Metadata Initiatives** - [Dublin Core Metadata Initiative](#), The [Music Ontology Specification](#), [RAMM.X](#), [ID3v2](#), and the [media-info](#) Microformat.

Finding a Community Standards Process that Works

The first thing that the company wanted to focus on was the problem of creating a standard way of representing music on the Web. There are several Web standards organizations that were considered. Among them were the Internet Engineering Task Force (IETF), The World Wide Web Consortium (W3C), OASIS, and a new community-based standards creation process called Microformats. The primary goal was to avoid any sort of red-tape and member-fees. There was uncertainty about what the standard was going to look like after it was created and the ability to make rapid changes in the first year or two was desirable. Those involved in the process knew that many mistakes were going to be made. Most importantly, Digital Bazaar desired input and adoption by a large number of web publishers.

The IETF, W3C and OASIS didn't make the cut due to the amount of red-tape and politics that were bandied about on forums and web sites. Digital Bazaar had no experience with any standards bodies but knew enough people that had nightmare experiences when creating a standard through a large standards body. The goal was to create a music markup standard in four to six months, and none of the larger organizations had such a standards track. The choice was a pragmatic one - being a start-up, the company could not afford a large round-trip on a standard. The company focused on pairing itself with a community that could innovate at a rapid pace and cared about website publishers first and foremost.

Web searches described a fairly new community forming around a concept called [Microformats](#). The community stated that they put publishers first, had a community-based standards creation process, and enabled the semantic web through a very simple mark-up mechanism that worked in

all versions of HTML. The Microformats [community](#) and [process](#) was where Digital Bazaar would take its first shot at a Web-based music markup standard.

Working with the Microformats Community

After more research, it was discovered that the Microformats process solved a number of long-standing company issues:

- Implementation of Microformats are [very simple, and publisher friendly](#)
- There was a [process](#) in place for creating new Microformats.
- The [community](#) was open to the general public.
- The Microformats Process was open to aggressive revision, allowing the music Microformat to change during the first year or two.
- Microformats focus on the practical and a scientific process for creating new standards.
- There were no politics, red-tape or cash investment needed to participate - it was an open community built around the concept of creating web standards.

Our employees were elated to find that such a community existed and that it had already started an initiative called [media-info](#) to semantically mark up video, audio and images on the Web. After some [discussion](#) on the Microformats new mailing list, the company decided to split off and take the initiative on creating the music part of the media-info project.

This was called the [hAudio Microformat](#) and work began in earnest. All contributions followed the Microformats process in order to identify a potential grammar that would solve many of the woes experienced by music website publishers.

Creating the hAudio Microformat

The Microformats process is a scientific process that requires one to gather examples, analyze those examples, publish trends and patterns and finally formalize a vocabulary.

For the audio vocabulary, Digital Bazaar employees and Microformats community members gathered over 185 [examples](#) and analyzed each website, looking for properties such as "artist", or "album name", or "publisher" in each page. All analysis was recorded by the community through the [audio-info-examples](#) page on the Microformats wiki. The community also examined the music [metadata file formats](#) that existed in early 2007. From all of this data, a number of [trends and patterns](#) were discovered resulting in a preliminary vocabulary based on real-world analysis of published music data.

This approach is an important one to recognize because it leads to very few theoretical debates on what is and what isn't useful. In short, if one can't find examples of data being published on the Web, there is no reason to create a tag for that data because there is nothing to mark up. An example of the prevention of a theoretical debate came about as the community started to discuss the *rating* tag. Without data to backup one's arguments, a theoretical debate about the property could continue for days if not weeks. However, after analyzing all of the music websites, our employees discovered that [only around 10% of websites publish ratings for songs or albums](#). This was clearly under the threshold for inclusion and thus a great deal of time was saved in the Microformats community by not having a theoretical debate on the usefulness of a ratings vocabulary term.

The process helped the community create a minimal vocabulary that was backed by hard data

called the [hAudio Microformat](#). As a by-product, a collection Microformat pattern was developed to help describe music albums as well called the [hAlbum Microformat](#), which was later folded into the hAudio Microformat. The majority of the research for hAudio was completed in 3 months. A first draft of the hAudio format was created in another month - impressive output for a community-driven standards process.

Problems Encountered with Microformats

Following the Microformats process was not without its problems and frustrations. Like any new community, some kinks are still being worked out. Some of those kinks include:

- **The Process** - Mixed interpretations of the Microformats process lead to arguments about the process. These arguments distract new Microformat authors from their work. For example, not everybody agrees on what constitutes "enough examples", or "proper interpretation of W3C standards". The process is constantly being refined, and in some cases re-written, which is frustrating to those following the process.
- **Politics and The Cabal** - There are several involved in the community that are [bitter about how the community goes about making decisions](#). Some of the founders of Microformats, sometimes called *The Cabal*, quite blatantly go against some of their own rules and at times it seems as if rules are being applied to *The Cabal* differently than the rest of the community.

While the above are being worked out and will probably be resolved in the long term, there were three things that concerned key Digital Bazaar developers and will probably never be resolved due to the philosophy employed by the Microformats community:

- **Microformats Implementation** - The Microformats implementation is simple by design, which is a double-edged sword. Doing advanced mixing and matching of various Microformats is difficult because Microformats are both [scope-less](#) and have [no namespaces](#). This means that if one were to list an album with an artist and a track inside that album which is performed by another artist - the Microformat won't be able to differentiate which artist created the album and which one performed the song. It is also nearly impossible to link an artist at the top of a web page with an album at the bottom of a web page if there is any other album listed on the web page. While these might seem like corner cases, they are far more common than they might seem at first.
- **Being a Visionary** - Microformats do not allow a trend-setter to add anything to them that isn't backed up by hard publishing data. The side effect of this is that there will always be a group of people that need to publish higher fidelity data in a standards-compliant way than is supported by the Microformat. Microformats are designed to solve 80% of the web's publishing problems - the other 20% of web publishers are left to fend for themselves.
- **Centralized Vocabulary Development** - What started out as a major benefit, which was the ability to go to one location to discuss and formulate an audio vocabulary, turned into a hindrance when the time came to take hAudio further than what the examples on the Web supported. While Microformats were a great start, and "paved the cow paths", they didn't allow Digital Bazaar to innovate. To do that, a decentralized vocabulary development approach was needed. Bitmunk needed to be backed by a semantic web technology that didn't require the approval of a 300+ developer community to proceed.

It is important to note that while the above three are issues that are problems for the company, the Microformats community solved one very important problem for us: It allowed the creation of a basic Audio Microformat in roughly four months that got us 80% of the way towards solving the web's music publishing problem. The down-side is that the hAudio Microformat still has not reached the

draft stage after 20 months of development due in part to Digital Bazaars focus on completing RDFa as well as stagnation caused by the Microformats process.

The World Wide Web Consortium and RDFa to the Rescue

One of the more concerning issues with the hAudio Microformat implementation was that it was nearly impossible to mix and match audio and video Microformats. This was partly due to the no namespacing, no-scoping problems mentioned earlier and also because vocabulary terms from other Microformats are re-used quite heavily between Microformats.

While the hAudio Microformat would solve the music mark up problem for a large number of website publishers, our company wanted to also ensure that others would be able to extend the hAudio format. Microformats can lead to a catch-22 situation. If there are not enough people publishing a property, such as rating, that property will never become a part of any Microformat. If the *rating* property is never included in a Microformat, publishers will be less likely to publish it.

A primary problem with Microformats is that the technical implementation feels hack-ish and does not scale. Most Microformats are good for the masses, but not for the visionaries. Luckily, an initiative was found through colleagues in [Creative Commons](#) and [Mozilla](#) called [RDFa](#).

After a preliminary phone call with the Chair of the RDF in XHTML Task Force (The RDFa Task Force), Digital Bazaar decided that it would be wise to help the W3C finish RDFa. The end-goal of RDFa supported a great deal of the Bitmunk Semantic Web requirements, which would enable us to achieve our corporate goals. The RDFa Task Force, headed by Ben Adida, was kind enough to invite us to participate based on the work we had performed in the Microformats community.

Manu Sporny started working with the Task Force as an Invited Expert to the W3C in late 2007. A large part of RDFa was already completed thanks to the many months of work by the Task Force members. Unlike the Microformats community, the Task Force was for W3C members and invited experts only. This had several positive benefits. Among the benefits were that all of the task force members were incredibly well versed in knowledge representation, RDF, and the underpinnings of the Web. Every member of the task force was courteous, polite, gifted with a razor sharp intellect and hawkishly focused. Most of the rumors about the W3C quickly evaporated as the weeks turned into months.

In October 2008, RDFa was published as the World Wide Web Consortium Recommendation. This meant that it was ready for developers and web publishers to use.

RDFa is a solidly-engineered, light-weight, technology solution for semantic data markup. It supports scoping and namespacing without adding a great deal of complexity as well as addressing a number of other issues associated with Microformats. It would also allow Digital Bazaar to improve the hAudio Microformat, providing extensions where developers saw a need to innovate. The Audio RDF Vocabulary, based heavily on the hAudio Microformat, was created by Digital Bazaar during the last several months of RDFa work. If two or more websites wanted to add one or two properties to the Audio RDF Vocabulary, they could do so with RDFa.

In the end, Digital Bazaar used a combination of the Microformats process to create a basic vocabulary for audio called the hAudio Microformat. The company implemented the Bitmunk website using hAudio and then re-implemented the semantic markup in the site using the Audio RDF Vocabulary and RDFa.

Developments as a By-Product of RDFa Task Force Participation

During Digital Bazaar's interaction with the W3C, there were several tools that were created to help RDFa adoption. The first was a test suite for RDFa called [Crazy Ivan](#). The suite was deployed as a web service and is used to ensure RDFa parser compliance with a set of test cases approved by the RDFa Task Force.

The second piece of software that Digital Bazaar developed was the [librdfa](#) software library. The software library provides a very compact and blazingly fast C-based RDFa parser. The library is provided under a number of open source licenses and is released with the hope that it will speed RDFa adoption. The library has already been integrated into the Redland suite of Semantic Web tools released by Dave Beckett.

RDFa is built upon the Resource Description Framework (RDF) and is not very useful without a set of domain-specific vocabularies. In the case of Bitmunk, vocabularies were needed to mark up audio, video and digital commerce opportunities. Preliminary research and development conducted in the Microformats community led our developers to be fairly confident in a minimal set of semantic data that was being published on the web. Converting this knowledge into an RDF vocabulary was a simple task resulting in the [Audio RDF Vocabulary](#). Further research into video publishing on the web and re-use of the media-info Microformat research helped to create a [Video RDF Vocabulary](#). Analysis and re-factoring of the Audio and Video RDF vocabularies generated a base vocabulary called the [Media RDF Vocabulary](#). A [Commerce RDF Vocabulary](#) was created based on the hAudio Microformat research and Digital Bazaar's experience building digital media commerce systems for the web.

In order to test the functionality of librdfa and demonstrate real-world applications based on RDFa, Digital Bazaar created a Firefox 3 plug-in for detecting and operating on semantic data expressed in RDFa. The plug-in is called [Fuzzbot](#) and is a visual tool used to debug RDFa data found in a web page. Currently, Fuzzbot is in development and the user experience is fairly rough but does demonstrate some novel uses for RDFa semantic data interaction. Two videos demonstrating Fuzzbot operating on [audio](#) and [video](#) data are available via YouTube.

Standardizing Web-based Digital Media Commerce and the Future

The following section outlines what Digital Bazaar hopes to achieve by applying Semantic Web Technology to various digital content industries.

Kick-starting Innovation by Standardizing Semantic Expression

The Microformats community helped Digital Bazaar kick-start a large-scale community-driven semantic web audio markup format. However, it was RDFa that allowed the semantic audio markup approach to meet the necessary technical and scalability requirements. RDFa allowed Digital Bazaar to establish a base standard for the digital content commerce industry.

Digital Bazaar believes that there are places for both Microformats and RDFa to co-exist on the web, each with its own purpose. Microformats are a good solution utilizing simple markup and when applied to well-defined problems. RDFa is the best technical solution, is socially scalable by allowing distributed vocabulary development, and helps innovation happen through experimentation. It was because RDFa allowed developers to do everything that Microformats did as well as provide a sound technical foundation for Digital Bazaar's commercial interests that the company has decided to focus our efforts on RDFa integration into the Bitmunk toolchain and customer-facing software.

Digital Media Commerce Use Cases Supported by RDFa

The first two generations of the Bitmunk software were implemented using a fairly heavy-weight Java application. This had the benefit of running on most major operating systems and the severe limitation of needing a very large Java Virtual Machine to run the application. The execution footprint of the system was close to 150MB-189MB. The adoption rate for the first and second generation Bitmunk software was greatly hindered because it lacked strong web browser integration and required the Java Runtime Environment. Our third generation software, which is a complete rewrite of our first and second generation software, is written in cross-platform C++, supports tight Web browser integration via plug-ins, and achieves a very small execution footprint. The execution footprint of the third generation Bitmunk software is closer to 2MB-8MB, which is in the range of operating on cell-phone hardware.

One of the Bitmunk network goals is to give someone that is browsing the Web the option of purchasing digital media when they hear about it on blogs, social networks and any other site mentioning a song, movie, television episode or book. To enable this functionality, the software needed to understand when each type of digital media was mentioned on a web page. It was out of this need that the Audio, Video, Media and Commerce RDF vocabularies were created. It was also because of this need that technology like RDFa was required. RDFa allows the Bitmunk software and other web publishers to mark the audio and video metadata up in a standards-compliant manner.

The ideal customer experience for Bitmunk is as follows:

1. Your web browser detects an audio album and artist on a web page and notifies you in a discrete, non-obtrusive manner.
2. You instruct your web browser that you would like to see more information and the current prices for the album, without interrupting your main browsing experience.
3. The album information, encoded using the Audio Vocabulary and RDFa, is lifted from the web page and processed by the Bitmunk plug-in.
4. The Bitmunk service recommends a number of online locations that are selling physical copies of the album. The service also recommends sites that are selling digital copies of the album, which include sellers on the Bitmunk peer-to-peer network.
5. Upon noticing that several sellers on the Bitmunk peer-to-peer network are offering the digital album at a discount, you purchase the album from the Bitmunk network.
6. If you like the album, you write about it on your blog or website. The Bitmunk software would generate a chunk of RDFa that you could then insert into your blog entry. This chunk of RDFa would contain machine-readable semantic data about the album, including a direct purchase link to your Bitmunk sales software.

The Bitmunk website already publishes over 1.3 million pages using RDFa and the Audio RDF Vocabulary. Digital Bazaar is currently working on tools and utilities to empower web publishers to publish music, movies and television using RDFa and the community-driven vocabularies that have been mentioned in this article. It is important to note that all of the technology that is being used to accomplish the scenario described above is not only open source, but also uses well documented, open standards.

Overall, this endeavor has been a very positive experience for the company. Digital Bazaar looks forward to continuing to work with both the W3C and the Microformats community to standardize technologies related to making open, standards-based, peer-to-peer digital content sales a reality on the Web.

Key Benefits of Using Semantic Web Technology

While there have been many benefits while using open source and open standards to innovate in our particular industry, there are broader lessons that can be learned by companies interested in using Semantic Web Technology to realize their corporate goals.

Key Semantic Web technology benefits include the following:

- RDF provides a solid foundation for knowledge representation.
- RDF vocabularies provide a means of distributed innovation - a cornerstone of the World Wide Web.
- RDFa enables both human-readable and machine-readable semantics to be expressed using a single document.
- Even providing a minimal amount of semantics in a web page exposes data for a myriad of known, and yet-to-be-discovered uses.
- RDF and RDFa allows companies to innovate in parallel, allowing the market to decide the most popular solution.

© Copyright 2008, [Digital Bazaar, Inc.](#)

This document is released under a [Creative Commons Attribution Non-Commercial Share-Alike 3.0](#) license.