Transcript

[00:00] [background music]

Derek Bruff: [00:10] Welcome to "Leading Lines," a podcast about Vanderbilt University. I'm your host, Derek Bruff, Director of the Vanderbilt Center for Teaching.

[00:13] In this podcast, we explore creative, intentional and effective uses of technology to enhance student learning — uses that point the way to the future of educational technology in college and university settings. In this episode, we feature an interview with Steve Baskauf, senior lecturer in biological sciences at Vanderbilt University.

[00:30] Steve coordinates the introductory biological sciences labs, trains and mentors the undergraduate and graduate student teaching assistants for those labs, and designs and assesses inquiry-based lab curricula.

[00:41] However, this interview focuses on another aspect of his work at Vanderbilt, biodiversity informatics. Steve has developed Bioimages, an online image database with over 10,000 annotated plant and ecosystem images, and he has created mobile-friendly tree tours of the Vanderbilt Campus.

[00:57] The interview was conducted by my colleague, Cliff Anderson, Associate University Librarian for Research and Learning. Cliff and Steve discuss the Semantic Web, Linked Data and the challenges and opportunities of creating and using machine-readable datasets.

[01:09] [background music]

Cliff Anderson: [01:11] Hi, this is Cliff Anderson with the Leading Lines podcast. I'm here with Steve Baskauf. Steve, why don't you introduce yourself and tell us a little bit about your role at Vanderbilt?

Steve Baskauf: [01:23] I'm a senior lecturer in the Department of Biological Sciences and my primary responsibility is running the Intro to Biological Sciences labs for majors. However, on the side, I also have a biological images database and repository called Bioimages that I run.

Cliff: [01:49] One of the things that Steve does here is also keep a database of the trees on campus. That, I think, is also connected with your Bioimages project.

[02:00] Why don't you talk, just to introduce people because they may not know Vanderbilt has this wonderful arboretum, how your work got started with the trees on campus?

Steve: [02:11] I've always loved the trees on Vanderbilt campus, pretty much like everyone else, and I was interested in developing online tools that would allow people to explore the arboretum.

[02:28] Probably 10 or 15 years ago, I started developing an online tree tour. Actually, it wasn't really online, it was supposed to be run on your laptop from a CD. I quickly discovered that that really didn't work very well because I tried walking around campus holding my laptop and it was a little bit heavy.

[02:55] Right as I was pondering that, the iPhone came out. I realized that this is actually the platform that we need for allowing people to explore the arboretum, so I transformed the online tree tour that I had made for laptops into one that was optimized for portable devices and I've been working on that.

[03:19] In some ways, it's integrated with the Bioimages project because the images of the trees are a part of the Bioimages collection, but there's a separate website for the arboretum, which is at vanderbilt.edu/trees if you want to check it out. There's content there that's specifically developed for the purpose of helping people to learn about the history and the diversity of the arboretum.

[03:51] The actual tree tours that you can take on your phone, which are also accessible from the arboretum website, are based on the technology that underlies Bioimages.

Cliff: [04:04] This is a good opportunity then to introduce Bioimages, which is an incredibly rich project that you've developed over the last few years based on some really innovative technologies. We'll talk in detail about what we mean by the Semantic Web, but why don't

you just introduce the project?

Steve: [04:26] Bioimages started off as an idea I had after I took a plant taxonomy class. I struggled with trying to key trees out and other plants using a book that had no pictures in it, only the technical terminology. I thought this would just be so much easier if I could actually see what these plant features looked like.

[04:49] I started off by taking a lot of detailed photographs of different parts of plants. Not just the whole tree, but what do the buds look like? What do the leaves look like? What do the edges of the leaves look like and so forth, with the idea that these images could be built into tools that would help people to learn better.

[05:14] Bruce Kirchoff, who is one of my collaborators, actually has used them to create some online tools to help people more effectively learn how to recognize plants by sight identification, as opposed to keying them out.

[05:33] Early on, one of the issues that I struggled with was how do you keep track of the information about the images? The technical term we would use for that is the metadata. How do you keep track of the metadata? How do you keep track of the images? How do you link that all together?

[05:55] I started asking questions trying to figure out what the best practices were for that. That was basically how I got sucked into the whole Semantic Web thing.

[06:06] Bioimages, if you go to the website, which is bioimages.vanderbilt.edu, it appears primarily to be an image database. That is a major component, the images are there, but underneath the images, there's an information layer that's not quite so obvious. That information layer is built on the principles of Linked Data and the Semantic Web.

Cliff: [06:37] Why don't we just define both these terms? They're related, but they're not synonymous. What do we mean, first, by Linked Data? Why don't we start there?

Steve: [06:51] Linked Data…actually Semantic Web comes a little earlier than that. I'll just…

Cliff: [06:59] We'll go with Semantic Web first, sure.

Steve: [07:02] We can flesh this out a little bit more in a moment. Tim Berners-Lee, who basically invented the Web and wrote the standards for how web pages communicate with each other and how they're linked and that sort of thing. Soon after he conceived this idea of a human readable and traversable web, he also came up with the idea of a machine traversable web.

[07:31] Everyone is familiar with Google. Google scrapes content from the Web and then has to try to figure out what it means. When you do a Google search for a particular thing, Google has to basically guess what the web pages are about so that they can present you with the information that you're looking for.

[07:55] Tim Berners-Lee's idea was that there would be this parallel web to the human readable World Wide Web, which he called the Semantic Web.

[08:06] In that Web, machines or software would be able to basically know what web pages meant and would be able to go from one web page to another and learn things in a machine readable way, in the same way that a human surfs the Web by clicking on hyperlinks.

[08:30] That's basically the idea of the Semantic Web. In a machine readable document, you would basically encode information in a way that a machine could understand and that machine could also draw inferences that maybe weren't directly stated, but that were implied or entailed by the information.

[08:54] Perhaps it was contained in two different web pages.

[08:56] Linked Data is a lower level implementation of this. The primary thing that Linked Data depends upon is the idea of URI as both an identifier and also a way of acquiring information. Most people are familiar with a URL. You might see a URL on an advertisement or something. If you type it in the bar at the top of your browser, it takes you to a web page.

[09:36] What Linked Data does is basically says that a URI, which is kind of a superset of URLs, is not only a way to access or to retrieve information about a thing, but it's actually an identifier for the thing itself. We're used to the idea that a web page has a URL, but something like a tree, that is a physical object, can have a URI.

[10:08] If you put that URI in a web browser, it doesn't give you the tree, but rather it denotes

that this URI stands for that tree. If you put that URI into a web browser, it will give you information or metadata about that tree, preferably in a machine-readable format, but also in a human-readable format.

[10:38] This idea that URIs can stand for things and that you can connect things to other things through using URIs is really what underlies the idea of Linked Data.

Cliff: [10:55] One of the things that Steve's been doing is helping us to lead a working group on the Semantic Web. What's been fascinating is that we've discovered many faculty and students are coming together with similar sorts of problems and in this working group, we're applying these principles to their datasets.

[11:17] What's really nice about that is we talked about the two cultures, science and the humanities, and yet the Semantic Web really is bridging both of these cultures. There's a very strong cultural heritage community. I think there's also a very strong community in lots of the natural sciences, perhaps most in biomedical, but maybe you could say more about where that community is. You're a specialist in that area.

Steve: [11:43] My professional involvement in this is in the biodiversity informatics group, so I work a lot with an organization called TDWG, which is a biodiversity information standards organization.

[12:04] My work there has to do with developing standards for how you describe things like museum specimens and organisms, where they're located, what their taxonomic identities are, so how do you describe those in a machine-readable way and then link them to other things.

[12:25] Even though it seems like this is quite different from something like a cultural heritage project, there are actually a lot of similarities because there are agents which would be things like people and organizations involved in both of them. Usually they have some kind of geographic component, so places. There's a time component.

[12:49] There's a literature component, so how you refer to a journal article by means of a DOI, which is an example of a URI that stands for a particular thing, which in this case should be a journal article or a book or something like that. There's a lot of common problems of how you connect these things that are found both in the scientific community and in the

cultural heritage community.

[13:24] One of the projects that we've been workshopping is Tracy Millers' Chinese temple data. If you look the issues involved in her project, which is locations and metadata about temples and images, it's really very parallel to the same issues that I have with trees that are found on different geographic locations that are documented by images.

[13:51] The sorts of solutions that you would use with Linked Data between that kind of cultural heritage project in a biodiversity informatics project, they're really very similar.

Cliff: [14:09] The connections between these projects, I think they're similar, but on the other hand, one of the things that's challenging for people coming into this area is to become familiar with the vocabularies that they can use.

[14:23] A typical project in digital scholarship, people have a dataset, they elect to use a database -- typically, it's a relational database -- they put their information into rows and columns, maybe they'll do a little bit of a connection between tables and then develop a website, which works really well for the intention that they designed it.

[14:45] It gets on the web and people can query the data, but what's lacking is the connections to other projects and there's a great deal of repetition of information. For example, if we're talking about a tree database, if you're going to develop your project that way, you'd have to import all the information about geographic locations into your tables and every other aspect of the information that you wanted to share.

[15:12] The payoff for this kind of project is that you only maintain what's really core to your project and instead of replicating information, you link out to other people who are maintaining what's core to their project and benefit from their constantly updating and curating that information.

[15:28] The challenge, though, for people coming into this area is they don't know those vocabularies. You have to actually become acquainted with different communities working with different standards and understand how they describe the objects and information that they're curating. That's probably the biggest challenge for someone new coming into this area.

Steve: [15:48] Yeah, I think that's true. In some cases, there are very rich vocabularies that are already developed. For example, we have been looking a lot at the Getty Thesaurus of Geographic Names, which is a massive database about places and their names.

[16:08] That's actually a resource that both the Chinese temple project and my biodiversity informatics project could draw from the same, basically, vocabulary of place names and ways of describing geographic subdivisions and things like that. If there is something like that that's already developed, then the challenge is linking your data to it. In some cases, there aren't really good vocabularies that you can use.

[16:42] In that case, the challenge is really a social challenge, which is how do you develop a consensus among the various parties that have a vested interest in using a vocabulary like that, how do you go through that community process to develop a consensus or even a standard vocabulary for describing things in a machine-readable way.

[17:11] I've been fairly heavily involved in that with TDWG, this organization that I work with, in trying to get a consensus on what terms are we going to use to describe different things, and how are we going to link things together and so on.

Cliff: [17:30] Let's turn to the use of the Semantic Web and Linked Data as an educational technology because one of the really promising possibilities here for graduate students is to become involved in those communities and to learn how the communities describe the objects that they're talking about, which is a really, I think, powerful way to become acquainted with the field.

Steve: [17:55] I think the first challenge is to figure out who are the people who are working on the same kind of issues as you are and do they already have the Semantic Web or Linked Data effort. As we've been working on this project, we have discovered resources that are relevant.

[18:20] For example, a group that's developing descriptions of periods of time, like Chinese dynasties, and groups that are developing gazetteers of historical locations that maybe don't even exist anymore and so to pin…and there's people working on archeology and pretty much all of the different sorts of fields within the humanities and natural sciences. There's probably a community that's already working on developing some kind of standards or consensus vocabularies, if you can find what they are.

[19:01] The payoff is that you…it's essentially the network effect, which is that if you connect your information to someone else's information, then you end up with something that's greater than the sum of the parts, which sounds impossible, but that's where the Semantic Web part of it comes in.

[19:30] We didn't really talk about that, but one of the layers that gets overlaid on top of Linked Data is that the descriptions that you have of things, which are done in little units called RDF triples, it's not just a method of, say, marking up your data. Those data actually are considered to have meaning.

[20:06] The Semantic Web basically indicates that you can combine different bits of information and those bits of information may entail other information that no one is actually directly asserted. If you combine data, there's a potential that a machine could essentially learn new information that wasn't stated by anyone. That's the promise of the Semantic Web.

[20:37] How effective that has happened is questionable. People have been basically dreaming about this for over 10 years. It hasn't developed at nearly the speed as a human readable web. I don't know if we want to talk about some reasons for that.

Cliff: [20:58] Why not? What are the downsides here, or the impediments, maybe it's better to say?

Steve: [21:03] There are probably two major reasons why the Semantic Web has not developed as fast as people thought it would. The first one has to do with what I talked about it the context of Linked Data.

[21:20] It's really important that once you describe or once you denote something with a URI, that URI shouldn't go away. Many people have had the experience where they bookmarked the URL and then someone decided, "Oh, I don't like the way we organized this." We reorganize our company, so we're changing all our URLs to reflect our company's reorganization and then basically that link gets broken.

[21:52] If you're going to use URIs to represent things, you have to carefully think out how you are going to keep those URIs stable over a very long period of time. If it's information that you control and is fairly well contained like, for example, the Getty Thesaurus is run by the Getty organization, so they can decide on what URIs they're going to use and keep those

stable.

[22:24] If you have things that are more like the community resource, then coming up with some consensus on how the community is going to keep those URIs stable is actually really challenging and there's a number of different approaches that people have taken.

[22:41] For example, DOIs, which are very widespread in the library community, is a very successful and widely accepted solution for the problem of keeping identifiers stable, but in my community, in the Biodiversity Informatics' community, there really is not that sort of consensus and that's probably the major reason why there hasn't been a lot of progress there.

[23:09] The other part of it has to do with talking a common language. We touched on this earlier when we were talking about standardized vocabularies. One of the principles of Linked Data in the Semantic Web is that anyone can say anything about any topic. That's true, you can say anything you want in machine-readable language. The question is, is any other machine going to understand what you said?

[23:43] If you don't have consensus set of properties that everyone in the community is going to use to describe the resources that they're interested in, then if you don't have a consensus, you end up with 10 people each describing the same kinds of objects in 10 different ways.

[24:07] That makes it impossible to aggregate those data into a combined dataset that would allow you to make the discoveries that I was talking about before where you learn new things because of combining other people's data.

[24:26] That's another problem that is going to have to be dealt with. Different communities are at different places in terms of their success in defining vocabularies and then also, something we haven't touched upon, is graph models, which is probably beyond the scope of this podcast, but I guess Suellen talked about that in an earlier one of your podcasts.

Cliff: [24:56] We did. We had a podcast on Neo4j and the graph databases.

Steve: [25:00] Maybe I should just refer you to the other podcast and say that developing the third piece besides consensus on how to handle URI identifiers, common vocabulary and then a common graph model within your domain of interest.

[25:20] Those are the three major issues that a community has to come to grips with before you can get the synergism that the people who envisioned the Semantic Web thought was going to happen fairly quickly.

Cliff: [25:38] If I were someone coming totally fresh to this Semantic Web, where do you think the best place is for me to get started…Let's say, I'm a student, I'm interested in learning what's going on in my field. Where would you recommend someone to look first?

Steve: [25:56] That's a good question. The W3C, which is the World Wide Web Consortium, has a number of documents about different parts of the Semantic Web, such as RDF, and usually they have fairly technical specifications, but they also generally have user's guide and those are usually fairly accessible with examples and things like that.

[26:26] I have to say it's not an easy thing to just go and learn about on your own. I mean, there are books and they often have examples in them, but in order to try out the examples and actually do something with what you are creating, there's a technological barrier in terms of installing the kinds of software that you would need to make it run.

[26:56] I would advise people to find a group of people who are doing this kind of thing and then get involved with them because especially if there are some technically minded people in the group who've already dealt with some of the issues of setting up a query-able endpoint and things like that, then it would be easier for you to try out or test out the things you're interested in in the context of what they've already developed.

Cliff: [27:27] Yeah, I would second that. This is of course the group, that I referred to in the beginning, we have here called the Semantic Lab Working Group, which brings together people from all fields across Vanderbilt. Anyone listening to this podcast who wants more information about how that group works, we have a website which I'll put in the show notes.

[27:49] We'd also be happy to talk to you about setting up a group at your own institution. I think developing the community aspect, because people come in with lots of different skills, some technological, some more on the metadata side, it's good to have librarians involved and you need subject specialists.

[28:05] It really is an interdisciplinary community that you need to develop to make this work at its maximum promise.

Steve: [28:13] I thought about this whole issue of how do you get somewhere with Linked Data and the Semantic Web. If you think back, Tim Berners-Lee at some point in time made the first web page.

[28:27] When he did that, there was only one and it was fairly useless until there were more than one that were connected together, so really involved developing the infrastructure of servers and the protocol for communicating information, but then people had to make web pages and had to link them together. Until that happened, there wasn't this kind of network effect.

[28:49] The same thing you could say about Twitter. There was a first tweet, but until there were more than one person on Twitter, that tweet didn't go anywhere. That's why the social aspect of this is so important.

[29:02] Bioimages is cool that it has this layer underneath it, but unless there's 10 or 100 or 1000 Bioimages that are talking the same kind of language and that are linked together, there really isn't much of a benefit of it over the more conventional technology. That's another reason why it's taken this technology a while to get off.

[29:32] There is a growing commitment to Linked Data and the Semantic Web. We're seen that in places like the Library of Congress and other organizations who basically…even the DOI organization which provides Linked Data service for all its identifiers.

[29:52] We see some rather large players making commitments to this and I think that eventually may result in making it easier for more people to participate and for the network effect to take off in the same way that it did in other commonly used things like the World Wide Web and Twitter.

Cliff: [30:14] Why don't we end our conversation there. We can talk a lot longer, but this is a great introduction to the Semantic Web and Linked Data and the uses it has on a university campus. Before I close, we always ask our guest, what is your favorite analogue educational technology?

Steve: [30:30] It would have to probably be a pencil and a piece of paper. I have students who have for assignments to draw things like a cloning vector plasmid. I see them struggling, trying to draw and label circles on their computer. I often just suggest, "Why don't you draw

it with a pencil on a piece of a paper, and then you could always take a picture of it with your phone?"

[30:59] There's still something to be said for a pencil and paper, although I have been trying to ban myself from pieces of paper because I have a bad habit of losing them. There is that downside.

Cliff: [31:13] Files get lost, too, from time to time. Listen, thanks a lot, Steve. I really appreciate it. Great to have you on.

Steve: [31:18] Okay, thanks.

[31:18] [background music]

Derek: [31:19] That was Steve Baskauf, senior lecturer in biological sciences, interviewed by Cliff Anderson from the Vanderbilt Library. In the show notes, you'll find links to the Bioimages project, the Vanderbilt Arboretum Tree Tours, and the library's Working Group on Linked Data and the Semantic Web.

[31:36] You'll also find a link to Steve's Instagram account, VU Trees, where he posts some pretty incredible photos of Vanderbilt's campus foliage. You can find those show notes on our website, leadinglinespod.com, and on our Twitter where our handle is @leadinglinespod.

[31:51] Leading Lines is produced by the Center for Teaching, the Vanderbilt Institute for Digital Learning, and the Office of Scholarly Communications, as well as the Associate Provost for Digital Learning. Look for new episodes the first and third Monday of each month, except next month. We're taking the first Monday in January off for the holidays.

[32:06] Our next episode will be available Monday, January 16th. I'm your host, Derek Bruff. Thanks for listening.

[32:08] [background music]

[32:11] [background music]