A Semantic Landscape of the Last.fm Music Folksonomy
Using a Self-Organizing Map

Joseph Biberstine, Russell J. Duhon, Katy Börner, Elisha Hardy, Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington, and André Skupin, San Diego State University

Data
Last.fm is an Internet radio site where users discover new music based on their listening history. They can annotate musical items such as artists and songs with arbitrary tags, ranging from "rock" or "jazz" to "seen live" or "songs I absolutely love".

The original data set, collected during the first half of 2009, was sampled by removing all but the 1,000 most frequently used tags (out of 281,818 total), then discarding items not annotated with one of those tags. This reduces the number of items from 1,393,559 to 1,088,761. After sampling, the average item has 6.8 tags (3.8 unique tags). The top ten tags include: rock, electronic, seen live, indie, alternative, pop, female vocalists, jazz, classic rock, experimental.

Self-Organizing Maps
A self-organizing map (SOM) is a form of artificial neural network that generates a low-dimensional geometric model from high-dimensional data. It transforms nearness in the input space to nearness in the map space. The map is a landscape made up of a grid of small neurons, each having some level of association with each of the tags in the tag space. The map is randomly seeded and then trained on a collection of data by repeatedly (1) finding the most similar neuron for a piece of training data and (2) pulling each neuron's tag associations closer to the matched piece of training data. Early in the training, the pull extends across most of the map. By the end, it affects only a tiny range around the most similar neuron.

Project
The landscape is a 180 by 180 grid of hexagonal neurons.

Regions on the landscape are labeled by the tags with which their constituent neurons associate most strongly.

Dark brown borders separate regions defined by each neuron's single strongest tag association; those regions are marked with black tags.

Dark blue tags identify regions defined by the second-strongest association of each neuron, and so on, as shown in the legend below.

The size and shape of each tag roughly reflects the extent of its region. The actual size and shape are somewhat flexible to guard against label overlap.

Note that tags can appear in multiple locations, especially if they have strong associations with multiple terms that are not themselves strongly associated. For instance, "alternative" shows up as a second-level region within "seen live", but also exists as a top-level region elsewhere on the map.

Acknowledgments
This work is funded by the Cyberinfrastructure for Network Science Center at Indiana University, the James S. McDonnell Foundation, and the National Science Foundation under grant SBE-0738111. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.