

Citation parsing made easy
code4lib 2008 / lightning talks

Erik Hetzner

California Digital Library

erik.hetzner@ucop.edu

Turn this:

Gerber, John P. Anton Pannekoek and the
Socialism of Workers' Self Emancipation,
1873-1960. Springer, 1990.

into this:

title: Anton Pannekeoek and the Socialism
of Worker's Self Emancipation, 1873-1960

author: Gerber, John P.

publisher: Springer

date: 1990

Harder than it seems

Can't use standard parsing techniques

Rules-based systems require tuning
for different styles

Statistical parsing methods can help

Hidden Markov models

Train on your data

> 75% field level F_1

Submitted to JCDL 2008

Erik Hetzner

California Digital Library

erik.hetzner@ucop.edu