



Figure 1.1: Structured and unstructured data on the web

actually a very large amount of structured and semi-structured data. The vast majority of web content is being generated from databases and content management systems containing carefully structured datasets. However, the often rich structure that is available in such datasets is almost completely lost in the process of publishing such structured data as human-readable Hypertext Markup Language (HTML) pages (see figure 1.1). A key insight is that we would have made major strides towards the vision of a more Semantic Web if only we could publish and interlink the underlying structured datasets (instead of just publishing and interlinking the HTML pages after much of the underlying structure has been lost).