## 2.1 开放获取资源元数据的特点

本文通过对专家遴选的134个国内开放获取期刊网站和15个国外开放获取仓储进行调研，梳理了当前开放获取资源元数据的特点主要有以下几点：

1. 开放获取资源元数据描述粒度细。《NSTL开放资源描述元数据标准规范》中对开放获取元数据的描述定义了7类元素集：开放资源集合、开放资源作品、责任者、机构、开放资源使用权益、获取来源、管理信息，而这7类元素集又包含了众多元数据字段，如题名、其他题名、关键词、英文关键词、摘要、英文摘要、作者、机构、唯一标识号、总页数、起止页、参考文献数量、年、卷、期、期刊名称、全文下载链接等。相比于其他网络资源，开放获取资源所需采集到元数据更细。
2. 开放获取资源元数据描述复杂。开放获取资源元数据的字段众多，而这些字段通常是以不同的形式展现在网页中，造成了开放获取资源元数据描述的复杂性。主要表现为部分元数据字段在网页中是按照单个字段进行展示，如文章的标题、摘要等是以一个标签单独展示。而部分元数据则是多个字段组合成一条文本信息进行展示，如文章的年卷期等是组合成了一个文本信息，再以一个标签展示。
3. 开放获取资源的网页结构多变。在对国内开放获取期刊进行元数据采集调研过程中发现，部分期刊网站的元数据展示页面，在不同时期所采用了不同的网页模板。在结构发生变化的开放获取资源网站中，一般会存在1-3套不等的网页模版。而其他网络资源，如电商平台、论坛等，通常只采用统一的网页模版。

## 2.2 开放获取资源元数据分类

开放获取资源的元数据是按照不同的组织形式封装在结构化的HTML标签对中，进而在网页中进行展示。通过对国内外开放获取资源元数据的调研和特点的总结，根据开放获取资源元数据在网页上的组织形式，将开放获取资源元数据分为：单一型元数据和组合型元数据。

单一型元数据是指网页的一个HTML标签对中只封装了一个字段信息的元数据，如文章的标题字段，一般在网页中只用一个标签对进行封装展示。此类元数据信息揭示简单明了、层次清晰。在对该类元数据进行采集时，只需要提取HTML标签对中的信息即可。

组合型元数据是指网页的一个HTML标签对中封装了多个字段信息的元数据，多个字段通常是组合成为一个文本信息进行展示，如文章所属期刊的年卷期字段通常是组合成一条文本在网页中进行展示。此类元数据信息揭示较复杂。在对该类元数据进行采集时，提取HTML标签对中的文本信息之后，还需要进一步对文本进行处理，进而得到各个元数据字段信息。