开放获取资源元数据采集方法探究[[1]](#footnote-1)\*

黄政 张学福 谢能付

中国农业科学院农业信息研究所 北京 10081

591701477@qq.com

摘要：针对开放获取资源的特殊性，探究开放获取资源元数据的采集方法，为开放获取资源的学术研究奠定基础。通过对国内外开放获取资源的调研，梳理出开放获取资源具有元数据描述粒度细、注重数据质量和元数据描述复杂等特点。通过文献调研分析，总结出当前开放获取资源采集方法主要分为：(1)基于接口协议的元数据资源采集方法；(2)基于网络载体描述的元数据资源采集方法。同时从易用性和准确性两个角度对国内外典型网页采集系统进行比较分析与评价。提出了基于描述载体检查机制的开放获取资源元数据采集框架，从理论上解决了当前基于网络载体描述的开放获取资源采集所面临的问题，并对开放获取资源元数据采集方法的研究进行了展望。

**关键字**：开放获取资源采集，网页信息抽取，网页采集器

Open Access Resource Metadata Extraction Method Research

Huang Zheng, Zhang Xuefu, Xie Nengfu

Agricultural Information Institute, Chinese Academy of Agricultural Sciences

**Abstract:** In view of the particularity of open access resources, this paper has explored the method of open access resources metadata extraction, which lays a foundation for the academic research of open access resources. Through the research of open access resources at home and abroad, this paper sorts out the open access resources metadata have some characteristics, such as metadata descript more detail, focus more on data quality and metadata description complex. Through literature investigation and analysis, this paper summed up the current open access resource extraction method is mainly divided into: (1) based on interface protocol of metadata resources extraction method; (2) based on description of the network resource metadata extraction method. At the same time, the typical web collection system at home and abroad is analyzed and evaluated from usability and accuracy, and summarizes the main problems existing in the extraction of open access resources based on network description. this paper put forward a new extraction method which based on description checking mechanism of open access resource metadata. This method solves the problems of current extraction method based on description of the network in theory. Finally, the progress research of open access resource metadata extraction method is discussed.

**Keywords**: Open Access Resource Extraction; Web Information Extraction; Web Information Extraction System

# 1.引言

2002年布达佩斯开放获取计划，首先对“开放获取”进行了定义，即可以在公共网络上被免费获取，允许任何用户对该文献的全文信息进行阅读、下载、复制、分发、打印、检索、超链接，支持爬行器收割并建立本地索引、用作软件的输入数据、用于其他任何法律允许的用途。而在使用这些文献时用户不存在财力、法律或技术上的障碍，只需在获取文献时保持其完整性，而对文献复制和发布的唯一限制，或者说版权在该领域的唯一作用就是给予作者控制其作品完整性及作品被正确理解和引用的权利[1]。现在提供开放获取服务的资源逐步增多。DOAJ上收录了来自136个国家的11000多个开放获取期刊，可检索的论文达到2280000多篇[2]，在中国科技期刊开放获取平台(COAJ)上收录的期刊为650个，可检索的论文数量接近142000篇[3]。随着开放资源数量和质量逐步提升，其价值越来越逼近主流商业学术资源。对于越来越多可以免费获取的学术资源，无论是学习交流还是二次研究都有非常重大的价值。诸如DOAJ之类的国内外开放获取期刊平台，已提供了论文的检索、链接等功能，基本上能满足学者们学习交流的目的。但这些开放获取期刊平台并不提供元数据的采集。对于开放获取资源的深层研究，如学科领域发展趋势研究、知识演化研究等，研究者需要自己寻找方法对元数据进行采集。对开放获取资源的采集方法的研究就是在这种背景下产生。

## 1.1 开放获取资源分类

随着互联网的发展，开放获取资源形成了不同类型。目前最为常见的是开放获取资源包

括开放获取仓储(Open Access Archives)和开放获取期刊(Open Access Journals)等两种类

型[4]。有学者从不同的角度将网络交互类资源，如微博、论坛等，也归于开放获取资源中[5]。本文认为交互类资源在进行采集时，机器很难辨别其学术价值。因此，为了使采集资源更具有学术价值，本文将开放获取资源定位为开放获取仓储和开放获取期刊两类。

开放获取仓储(Open Access Archives)是遵循开放获取原则的一种电子版科技文献存储和检索的数据库。研究学者们通过向开放获取仓储提交已发表或未发表的研究成果，使得成果能迅速被其他学者学习并讨论，从而促进当前学科的发展。开放获取仓储中可以包括：电子文档、论文、课程资料、数据文件、声像文件、机构记录以及任何类型的数字文档。开放获取仓储主要分为两种类型：一种是收集某一特定学科或主题的学科知识库，另一种是由多个学术机构共同组成的机构知识库[6]。

开放获取期刊(Open Access Journals)是遵循开放获取原则的期刊。开放获取期刊经过同行评审，并通过网络载体，对期刊资源进行描述。开放获取期刊与传统期刊的区别不在于期刊的载体是纸本还是电子，而在于对期刊的访问方式和访问权限。所有的用户可以不受价格和权限的限制免费从互联网上获取此类期刊的学术信息[7]。

## 1.2 开放获取资源的特点

本文通过前期对国内外多个期刊网页和国外期刊平台的调研发现，开放获取资源比其他网络资源，如新闻等，有一定的区别。主要表现为：

1. 开放获取资源的描述粒度更细。《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。而一般的网络信息的描述，如新闻，只需要采集标题、正文、作者、时间等少量的几个元数据字段便可满足需求。
2. 开放获取资源更加注重元数据的质量。开放获取的元数据除了采集和发布之外，还要作为其他研究的基础，如知识演化分析、学科热点分析等都需要论文的关键词等元数据作为数据基础。开放获取资源的元数据的质量直接影响到其他研究的效果。因此相比于其他只进行采集和发布而不进行内容研究的网络信息，开放获取资源对元数据的质量要求更高。
3. 开放获取资源元数据描述更加复杂。开放获取资源元数据描述复杂主要表现在两个方面：一是网页中元数据展示多样，即不同开放获取资源的平台网站和期刊网页所揭示的元数据薄厚不均，二是描述元数据的网页结构多变，如在一个期刊网页中，存在网站建立初期所采取的网页模板与现在不一致的情况。而一般的网络信息，如论坛信息，某一个论坛网站在展示信息时采用的是同一个网页模板，不存在不同时期网页结构不一致的情况。因此开放获取资源元数据描述更加复杂。

# 2.开放获取资源采集方法

开放获取资源主要分为两种：开放获取仓储(Open Access Archives)和开放获取期刊(Open Access Journals)。开放获取仓储(Open Access Archives)一般有数据提供方和服务提供方两种角色[8]。两者之间遵循统一的接口协议，数据提供方负责数据按照规定的格式进行存储，以供服务提供方调用统一的接口进行元数据的采集。开放获取期刊(Open Access Journals)是遵循开放获取原则，通常以Web页面为载体来描述元数据的期刊。用户通过对Web页面的访问来获得开放获取资源的元数据。而不同的期刊在通过Web页面对元数据进行描述时，并没有统一的规则，因此无法形成统一的数据采集接口。因此开放获取期刊需结合网页信息采集技术和开放获取资源的特点，完成采集工作。

开放获取仓储中，服务提供方通过接口协议对数据提供方提供的元数据进行采集的方法，本文称为基于接口协议的元数据采集方法。而对开放获取期刊，采用网页信息采集技术对期刊描述载体中的元数据进行采集的方法，本文称为基于网络载体描述的元数据采集方法。基于接口协议的元数据采集方法适用于相互合作的机构之间，采集效率高，但灵活性不强。基于载体描述的元数据采集方法适用于普通用户，同样也可以对开放获取仓储的网页元数据进行采集。在制定好采集策略后，系统开发人员开发出采集系统，便可进行元数据采集和利用。下面就这两种方法的实现方式进行具体的讨论。

## 2.1基于接口协议的元数据资源采集方法

此类元数据资源采集的特点在于数据提供方都遵守某些特定的元数据收割协议，供数据采集方对元数据进行采集。其中具有代表性的是OAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting)协议接口[9]。OAI协议于2001年1月推出了Version 1.0版，同年的7月推出了OAI协议的Version 1.1版。目前，OAI协议最新的版本是2002年6月发布的2.0版。

基于OAI协议的元数据获取框架主要是由三部分组成：数据提供方(Data Provider)、服务提供方(Service Provider)以及注册服务器（Registrar Server），如图1所示。

数据提供方负责元数据的生成，即按照统一的规范，在本地建立本地元数据仓库（Repository），并且以OAI的响应(response)向服务提供方发布元数据。

服务提供方通过OAI请求(request)把数据提供方发布的元数据进行收割，并对元数据进行整理加工，为用户提供统一的检索界面，并提供增值服务。

注册服务器是对服务提供方和数据提供方进行管理的模块。数据提供方与服务提供方都需要在注册服务器进行注册。

Repository

Repository

Repository

Data Provider

Service Provider

OAI Verb

Record

Request

Response

Registrar Server

图 1 OAI元数据获取框架

OAI指令是数据提供者与服务提供者之间的基于HTTP协议的应用通讯协议，主要由6个请求（request）和相应（response）指令。

1. Identify 用于获取元数据信息仓储的相关信息;
2. ListMetadataFormat 用来获取元数据信息仓储所支持的元数据格式;
3. ListSets 用来得到信息仓储集的结构,获取集名称等;
4. ListIdentifiers 请求获取可以从元数据信息仓储中查到的记录的标识符;
5. ListRecords 用来抓取一个数据仓储中的多条完整的元数据记录;
6. GetRecord 从一个元数据信息仓储中抓取一条具体的元数据记录。

数据提供方与服务提供方的一般交互过程如下：

1. 服务提供方向数据提供者发起一个HTTP请求；
2. 数据提供方对此HTTP请求进行解析，判断是否为OAI-PMH协议中的六个有效请求之一；
3. 若为有效请求，数据提供方根据请求进行相应，返回元数据结果，若不是有效请求则返回错误信息。

目前在OAI官方网站登记的OAI仓储有2997个[10]，基于OAI-PMH数据源的元数据采集技术已经十分成熟。许多学者也对基于OAI-PMH协议的元数据采集做了大量的研究。文献[11]在总结元数据的采集策略时提出，在抽取基于OAI-PMH协议的元数据时，通常首先用OAI协议的ListMetadataFormats命令查看该数据源提供的可选元数据格式，便于设定查询参数metadataPrefix的值然后用ListRecords、from(资源创建时间的上限)、until(资源创建时间的下限)和metadataPrefix的参数及值共同构成采集资源的入口地址同时可根据需求更改时间参数from和until的值，用于对已有采集数据的监测和更新。文献[12]利用OAI-PMH协议的特点以及OAI-PMH技术构架，构建了基础医学信息共享系统，将数据提供者和服务提供者，都作为OAI-PMH架构的一部分，从数据提供和数据检索都遵循OAI-PMH协议，解决了各个数据库存在数据格式不统一、管理不规范等问题，实现了础医学科学领域中数据的共享。文献[13]教授通过将OAI仓库中每一条记录分配一个URI，并用现有的以及自定义的词汇表和元数据集对URI对应的仓储所具有的元数据记录进行描述，再采用不同的方式将描述的关联信息转化为关联数据，完成OAI-PMH元数据的关联数据转化，使用户通过URI直接访问元数据记录，也可以通过SPARQL协议任意设定查询条件来实现对元数据的批量检索，从而避免用户在采集OAI仓储中的信息时所遇到的不便。

基于协议的元数据采集特点在于：数据提供方和服务提供之间采用了统一的协议，提高了元数据采集收割的效率。适用于遵守OAI-PMH协议采取统一的数据存储规范并通过注册形成的机构群体。但不足之处在于对数据格式要求十分严格，不具有灵活性，且对于网络上存在的大量的非统一规范的数据来源，采集十分困难。

## 2.2基于网络载体描述的元数据资源采集方法

随着互联网技术的发展，越来越多的学术资源在元数据描述上不再仅仅局限于纸质载体，而更多的是利用互联网，通过将元数据进行加工，形成电子文档，并以Web页面为载体来展示。近年来越来越多的学术资源加入到开放获取的行列中，这也使得元数据的描述更加多样化。资源提供方将存在本地数据库中的元数据按照自定义的描述规则，通过网页载体进行描述。而元数据的采集方对网页进行解析，编写抽取规则来抽取其中的元数据，并将抽取到的元数据保存在本地供研究使用。

Web信息抽取的定义描述为[14]：对于给定的一组Web页面S，定义映射W，W将S中的对象映射到一个具有更为结构化、语义更为清晰的数据结构SD(如XML、关系数据库等)，并且映射W对与S在语义和结构都相似的Web页面集合S'具有相同的功能，其中映射W就是Wrapper。

### 2.2.1 Web信息采集方法分类

Web信息采集的研究很早就开始了，并出现了许多网页采集系统。从不同的角度可以将Web信息采集方法分为不同种类[15]。不同的分类间会有重复。根据网页信息源的不同可以分为自由文本采集、半结构化的文本采集和结构化的文本采集。根据自动化程度可以分为人工方式、半自动方式以及全自动化方式[16]。根据包装器原理的不同可以分为基于层次结构的信息抽取归纳方法和基于概念模型的多记录信息抽取方法。根据人工干预的程度分为基于知识工程的方法和自动训练法[17][18]。根据用户和抽取系统的交互将网页信息采集系统分为手工构造的信息抽取系统、基于监督的信息抽取系统、基于半监督的信息抽取系统、无监督的信息抽取系统[19]。但最常见的分类方式是根据信息采集工具所采用的采集原理和采集方式的不同将网页信息采集分为基于自然语言处理的网页信息抽取、基于本体（Ontology）的网页信息抽取、基于包装器归纳方式的信息抽取、基于HTML页面结构分析的网页信息抽取、基于Web查询的网页信息抽取和基于页面视觉特征的网页信息抽取[20][21][22][23][24][25]。

基于自然语言处理的网页信息抽取是较早出现的信息抽取方法，在很大程度上依赖于自然语言处理技术(NLP,Nature Language Processing)。通过将内容分割成多个句子，分析句子成分，使用包括过滤、词性标注在内的自然语言处理技术来建立短语和短语之间的关系，进而获得抽取规则，所产生的抽取规则通常基于语法约束和语义约束[26]，主要适用于含有大量文本且句子完整、适合语法分析的开放获取资源页面。如基于WHISK[27]算法的信息抽取系统是基于自然语言处理的信息抽取技术的代表。

将此方法运用于开放获取资源采集时，是将网页文档当作文本进行处理。需先根据开放获取资源的网页样本进行标注，建立语言字典，再使用机器学习拟合抽取规则，构建抽取包装器。虽然许多学者通过改进机器学习的方法来提高抽取规则拟合的速度和准确性，但依然需要前期准备大量的训练样本和构建足够完全的语意字典。而且不同资源平台的网页结构不同，无法形成统一的抽取规则，导致工作重复且工作量大。

基于本体(Ontology)的网页信息抽取是一种针对特点主题或领域的网页信息抽取方法，是利用对数据本身的描述信息实现抽取。即在网页信息采集前，选择相关本体构建工具，根据选定的特定主题或领域，按照本体构建的原则，建立该主题或领域下的本体集。然后通过网络爬虫工具，对网页进行解析，完成对网页正文信息提取、分词等。将处理后的网页信息传递给计算模块，计算模块的主要功能是根据本体集来对网页信息进行加权计算。权值反映的是该网页与主题或领域之间的相关度，若与内容相关性高就保存，不相关就丢弃。中南大学许德智教授的课题组研发的SNAX[28]系统就是典型的基于本体的网页信息采集系统

在将此方法运用于开放获取资源采集时，需先请领域专家对不同来源的开放获取资源构造完整的本体知识库，此步骤工作量大，需要专家投入相当多的时间。在本体模型建立之后，就可以较少依赖于网页的结构信息，根据本体知识库中的信息对网页进行处理,数据的抽取就可以实现自动化，并且具有很好的适应性[17,21]。此外此方法对多种网页可以生成多种抽取规则，弥补了基于包装器网页信息抽取只能抽取单个网页结构的不足。

基于包装器归纳方式的网页信息抽取，即根据事先由用户标记的样本实例，应用机器学习方式归纳算法，生成基于定界符的抽取规则。定界符实质上是对感兴趣信息语义项的上下文描述，即根据语义项的左右边界来定位语义项。通常一个包装器只能处理一种特定的信息源[20]。如WIEN[29]系统。

此方法运用于开放获取资源时，可以解决具有相同结构和编码页面的信息采集，而且所花费的工作量较其他几种方法而言要少一些。但是对存在结构变化的开放获取资源网站采集效果不好。用户无法提前知道从那个阶段网页开始变化，因此无法选择比较全面的样本，形成的抽取规则也无法采集整个网站中所有的开放获取资源。

基于HTML页面结构分析的网页信息抽取，是充分利用了HTML的半结构化特性，在信息抽取之前通过解析器将Web文档解析成结构树，通过自动或半自动的方式产生抽取规则，定位主题信息，将网页信息抽取转化成对解析树的操作[30]。此方法将多个HTML文件作比较，找出其相似特征和不同特征，并基于这些特征生成由正则表达式所表示的包装器，再根据这个的包装器去抽取同一个网站中新的网页。RoadRunner是典型的采用基于HTML页面结构分析的网页信息抽取系统[31]。

在开放获取资源的网页中，很多是对数据库中数据的映射，是通过一个相同或者类似的模版所生成的页面，有一定的规律性。通过此方法可以对采用相似模板的开放获取资源进行有效的采集。但在形成抽取规则时，为了归纳出通用的正则表达式，要使用了大量复杂的启发式搜索算法，实现起来有一定的难度。

基于Web查询的网页信息抽取，更加深入的利用了HTML半结构化的特征，充分利用HTML的语法结构，将Web信息抽取转化为使用类似于SQL语句的Web查询语言对Web文档的查询。具体而言是将Web文档解析成一棵抽象的HTML语法树，然后在信息抽取之前，需根据页面的结构和所需提取信息的唯一的标记，写出合适的查询语句实现信息抽取。基于Web查询的网页信息抽取属于较为简单的抽取技术。Web-OQL[32]就是典型的基于Web查询的网页信息抽取系统。

在将此方法应用于放获取资源的采集时，由于封装元数据的HTML标签都具有唯一、特定的标识，需要采集人员对网页结构进行人工识别，如文章标题的HTML标签的id会包含“title”关键字。因此将此方法运用与开放获取资源的采集时，可以直接定位到需要抽取的对象上，因此抽取的元数据的准确性很高。不足之处是需要人工对网页进行识别，且对网页结构依赖性很强，对没有明确的唯一标识的网页抽取十分困难。

基于页面视觉特征的网页信息抽取，是在使用包装器对信息进行采集之前，利用Web信息的视觉特征，例如字体的大小和颜色、段落的长短、数据所处区域等信息，将一个网页分割成多个不同的视觉信息块，去除掉诸如导航条、广告条、版权信息等与主题无关的信息之后，再使用包装器对信息进行抽取，潜在的提高了包装器采集网页信息的准确性。基于VIPS[33]算法的采集系统辨识典型的基于页面视觉特征的网页信息抽取系统。

此方法在运用于开放获取资源采集时，虽然可以提高包装器抽取的准确度和效率，然而由于视觉特征具有复杂而又不确定的特性，往往需要人工不断地修改调整抽取规则，使用起来很难保证规则集的一致性[34]。

总结而言，基于不同方法的网页信息采集，最终都是使用抽取规则对网页信息进行采集，区别在于形成抽取规则的方式不同。基于自然语言处理的信息抽取和基于本体(Ontology)的网页信息抽取事先需要领域专家的参与，编写语义字典或本体知识库，且专家编写的质量直接影响到采集效果。此类方法更加适用于对网页中相关主题信息的提取，但因受限制于专家的参与，通用性较差。基于Web查询的网页信息抽取和基于HTML页面结构分析的网页信息抽取需要用户对网页结构进行一定的分析，对单个或结构统一的网页元数据采集十分方便，但对于向开放获取资源这样结构多样的网页，就需要多次人工解析，工作量大。基于页面视觉特征的网页信息抽取适用于结构清晰的网页，而且往往还需结合其他方法对元数据进行采集，适用范围有限。基于包装器归纳方式的信息抽取通过样本页面归纳学习抽取规则，减少人工参与，而且适用于各种类型的网页。但是在样本网页的选择中有时无法覆盖到所有情况，如在开放获取资源的网站中，无法预知从哪一个网页开始结构发生了变化，训练所得到抽取规则无法抽取网页结构变化后的元数据信息。而各采集方法通过不同方式最终形成的抽取规则，采集到的元数据都有较高的准确率和召回率。因此说明这些方法在其适应的领域都能保证采集数据的质量。

### 2.2.2 Web信息采集系统比较

当前Web信息资源采集方法，应用到开放获取资源采集中或多或少存在一些不适用性，而且大多数系统并未得到广泛的使用。为改进当前Web信息采集方法，使其更适用于开放获取资源元数据的采集，本文对当前国内外几个典型的网页信息采集器进行比较和分析。

表 1国内外典型的网页信息采集器比较和分析

|  |  |  |  |
| --- | --- | --- | --- |
|  | 特点分析 | 优点 | 开放获取资源采集中的不足 |
| 八爪鱼采集器[35] | 八爪鱼采集器通过内嵌的浏览器与用户进行交互，用户在内嵌浏览器中点击所需爬取的信息，系统记录所选取信息在网页中的唯一标识，并通过此标识来提取此系列其他网页中的信息。 | 简单的可视化交互，无需用户分析网页结构。可对不同深度的网页进行循环抽取。可进行云采集但需要付费，可设置定时采集。 | 网页中唯一标识发生变化，则无法提取信息。 |
| 火车采集器[36] | 火车采集器依赖于用户对网页的解析，通过用户查看网页源代码，找到所需提取信息的唯一标识，或者找到能分割采集区域的标签字符串，或者根据所提取信息的网页标签中的规律编写正则表达式来生成抽取规则。 | 可通过正则表达式扩充元数据的抽取规则，来应对网页中唯一标识变化的情况，提高抽取的准确度。支持丰富的拓展。 | 非常依赖用户对网页结构的识别，对正则表达式要求较高，不能定时采集，不如八爪鱼采集器易操作。 |
| ET采集器[37] | ET采集器是通过用户对网页源代码进行分析，人工识别需要提取信息的网页源代码，将不变的区域保留，将变化的区域设置为变量标记，以此来生成类似正则表达式的抽取规则。通过内置的测试工具能快速的查看当前抽取规则的抽取结果，并反复调整抽取规则，来获取最佳的抽取规则。 | 通过自动生成类似于正则表达式的抽取规则可提高抽取的准确度，内置的测试工具可以快速的查看抽取结果，并对抽取规则进行调整。 | 需要用户反复对网页源代码进行分析。 |
| 网络矿工采集器[38] | 网络矿工采集器将用户所需采集的网页分为导航页和采集页，网页URL可以通过系统所提供的按钮设置参数。对需要提取的信息，系统需要用户通过对网页源码的解析，找到所需要提取信息所在的页面代码块，通过设置所需提取信息的起始位置和终止位置，来设定类似正则表达式的抽取规则。 | 类似正则表达式的抽取规则使采集准确度高，系统提供大量的按钮和提示，减少了用户使用的难度。 | 需要用户通过识别网页结构寻找抽取信息的起始位置。 |
| Import.io[39] | Import.io是国外一款较流行的网页信息采集系统，采用的是B/S架构。用户通过输入URL，系统自动对网页中的信息进行采集。用户也可以通过简单的点击操作，选择其他想要抓取的信息。然后通过选择保存数据的格式，将采集到的数据保存在本地。 | 交互简单，易操作，对用户要求非常低，工作量少，适用于列表类的网页。 | 多数开放获取资源的元数据页是非列表类型，采集准确度不高。 |

国内外网页信息采集工具基本上是采用的是通过页面结构解析，归纳形成包装器的网页信息采集技术，主要是针对结构化很强的网页来进行信息的抽取，抽取过程中不需要大量的训练样本，只需用户对网页进行初步的分析，定位所需抽取的信息，自动或手动的生产抽取规则。

从易用性的角度来看，Import.io和八爪鱼采集器提供内嵌的浏览器，用户只需在浏览器中点击所需采集的信息就能自动生成抽取规则。而其他采集器都需要用户分析网页源代码，因此易用性上，Import.io和八爪鱼采集器最优。在需要用户分析网页的采集器中，网络矿工采集器提供了大量的按钮与提示信息供用户通过点击生成抽取规则，减少的用户使用的难度。ET采集器提供的测试工具可以使用户快速的查询到当前抽取规则的抽取结构，进而调整抽取规则，在一定程度上也提高的了系统的易用性。因此网络矿工采集器和ET采集器在易用性上次之。火车采集器提供了多种方式来生成抽取规则，如前后截取、正则提取、XPath提取、JSON提取等，有些方式需要用户有一定的专业知识背景才能生成有效的抽取规则，因此对于普通用户的使用有些难度。

从准确性的角度来看，本文通过对某一期刊网站进行试验发现，对于单个页面或者多个具有相同编码特征时，如文章标题的HTML标签中的唯一标识都为“J\_biaoti”，经过多次调整采集规则，所有采集器采集都能准确采集。但当此期刊网页中的编码特征变化，如文章标题的HTML标签中的唯一标识由原来的“J\_biaoti”变成了“J\_ZhaiYao\_BiaoTi”，各采集器都无法采集到相关信息。因此采集器的准确性除了与采集规则有关之外，与网页结构的统一性也有非常大的关系。当网页结构统一，通过调整采集规则，各采集器的准确率会非常高，但是网页结构不统一，则需对不统一的网页重新编写采集规则。

通过以上分析得出，各类采集器基本上能适用于网页结构一致的开发获取资源。但对于多种网页结构的开放获取资源，则需要编写多个采集规则。而且由于采集器在采集开放获取资源信息时，只能采集标签对中的信息，而多个元数据信息可能存在于一个标签对中，因此还需要对采集器采集到的文本进行分割、删减等二次处理，才能得到最终需要的元数据。

# 3.开放获取资源采集框架

基于接口协议的元数据采集方法在开放获取资源的采集中已十分成熟。机构之间通过规定的接口协议完成元数据的存储与采集。但此种方法受限于资源的存储必须遵循相关协议，使得方法的使用范围只局限于遵循协议的机构，而当前基于网络载体描述的开放获取资源占据了开放获取资源的大部分，此类资源载体描述的方式多样，当前Web信息采集方法和采集系统满足不了资源元数据的采集。通过对不同的Web信息采集方法在开放获取资源采集上中应用分析，以及从易用性和准确性两个角度对主要网页信息采集系统在开放获取资源采集上的对比，可以看出当前基于网络载体描述的开放获取资源采集存在以下几个共同的问题：

1. 采集系统对开放获取资源采集的通用性较差。当前国内外的网页采集系统中只针对了一个或者部分相同结构的网页进行元数据抽取。而开放获取资源元数据通过网页载体进行描述的方式多样，某一资源平台中存在多种网页结构的现象。而当前的采集系统只能根据抽取规则提取某一种网页结构中的元数据。其他网页结构只能在一次抽取结束之后再多次重新编写抽取规则，使得工作重复和降低工作效率。
2. 采集方法对开放获取资源采集的自动化程度不高。由于大部分信息抽取方法对开放获取资源的采集，都需要一定复杂的人工参与。如基于HTML页面结构分析的网页信息抽取和基于Web查询的网页信息抽取需要人工识别网页结构，基于自然语言处理的网页信息抽取前期需要构建语义字典，基于本体的网页信息抽取则需要专家在前期手工建模构建本体库等。各采集方法需要人工较复杂的参与，自动化程度不高，工作量较大。
3. 开放获取资源采集的准确性与效率较低。由于开放获取资源采集的元数据字段比其他网页信息抽取(如新闻、论坛等)多，且结构多样，往往对一般网页信息抽取表现较好的方法或系统，在抽取开放获取资源的元数据时准确率和效率就有所降低。

针对以上几点问题，本文结合采集方法和采集系统的特点，在基于包装器归纳法的基础之上，结合八爪鱼采集器的技术特点，提出基于描述载体检查机制的开放获取资源元数据采集框架，具体如下图所示：

用户点击需要采集元数据

开放获取期刊/仓储链接

系统自动形成抽取规则

信息抽取

用户交互模块

抽取规则库

检测URL和抽取规则是否已经存在

否

循环检测当前网站其他网页结构是否变化

增加当前抽取规则到规则库

否

元数据存储

数据库

检测模块

元数据采集与存储模块

是

是

图 2基于网页结构检查机制的开放获取资源采集框架

采集框架主要分为三个模块：用户交互模块、检测模块和元数据采集与存储模块。

用户交互模块，类似于八爪鱼采集器，通过内嵌浏览器与用户进行交互。在此模块中用户在内嵌浏览器的网页中点击所需提取字段，无需人工解析网页源代码，后台自动形成抽取规则。不仅降低了对用户的要求和工作量，而且通过用户自主的选择所需提取的字段，使得采集信息的准确性和可拓展性都得以保证，而且使用于任何形式的网页结构，通用性得到增强。

检测模块分为采集规则检测和网页结构检测两块。采集规则检测主要是对当前采集的URL和抽取规则在抽取规则库中进行检查，如果存在，则证明该资源已经采集过，就直接从数据库中将之前采集到的元数据根据用户需求返回给用户。如果不存在，则说明该资源未经过采集，将当前采集，就进入到网页结构检测。网页结构检测是对网站中所有需要采集的网页进行循环检测，查看是否有网页结构发生变化。若没有就进行网页元数据的采集。若有变化则再进入用户交互模块，让用户重新选择，形成新的采集规则，再进入检测模块。此模块一方面解决了开放获取资源网页结构变化的问题，提高了采集的覆盖度，另一方面对已采集到数据不用再进行重复采集，使得资源能重复利用。

元数据采集与存储模块是指根据当前的采集规则库中的采集规则，使用包装器归纳法对当前资源进行采集。采用类似于当前较流行的采集系统中所使用的分布式采集和存储策略，提高采集和存储的速度。

此框架一方面基于传统的采集方法，使元数据采集的质量得到保证。另一方面结合了当前流行的采集系统与用户交互的特点，在减少了一定工作量的同时也降低了对用户的要求。而最重要的是增加了检测模块，通过自动检测采集网站的网页页面结构和提取记录，解决了开放获取资源结构变化多样的问题，提高了开放获取资源采集的效率。

# 4.总结

随着开放资源的逐渐增多，以及数据挖掘、机器学习等技术在实际场景中运用的更加成熟，使得对开放获取资源元数据的利用价值更加丰富，如知识演化分析、学科热点分析等。开发获取资源元数据采集的研究是众多研究的基础，元数据采集的质量和效率将直接影响上层研究。因此对开放获取资源元数据采集的研究具有重要意义。通过对开放获取资源的特点进行调研，以及对现有采集方法和系统存在的问题进行分析，发现开放获取资源采集当期所面临的主要问题是采集系统的通用性较差、采集方法自动化程度不高、抽取准确性与效率较低。因为文本结合了采集方法和采集系统各自的优点，提出基于描述载体检查机制的开放获取资源采集框架。此框架从一定程度上解决了当前开放获取资源采集的问题。

最后本文对开放获取资源采集做了几点展望：

1. 普及统一的开放获取资源共享协议。虽然OAI-PMH协议已经提出很久，但只有少部分的OA仓储提供OAI-PMH接口共元数据采集，大部分的开放获取资源并没有提供统一的元数据收割协议接口，而且每个平台提供的元数据薄厚程度不一、页面揭示多样，使得开放获取资源采集比较困难。随着开放获取运动的不断深入，开放获取资源的不断增加，需要将统一开放获取资源共享协议运用到所有开放获取资源平台中，提供相应的采集接口，使信息的传递更加方便、快捷。
2. 多种采集方法集成使用，借助机器学习算法提高采集的效率。由于基于网页的开放获取资源具有多样性的特征，单一的采集方法无法全面的覆盖所有的网页，因此可以多种采集方法集成使用，并辅以智能的算法，提高大数据量的资源的采集效率。
3. 元数据采集趋于自动化、系统化和可拓展。随着开放获取运动的逐步深入，开放获取资源的分类逐步清晰，对每一类资源的特点也更加明确。因此针对每一类开放获取资源都会制定对应的采集策略，进而整合称为系统化的采集工具。对于新增加的开放获取资源类型，只需增量的添加到系统中，实现开放获取资源采集的自动化、系统化和可拓展。

总而言之，由于开放获取资源来源的多样性和复杂性，使得开放获取资源元数据采集需要更多的专业学者参与到研究当中。积极参与开放获取共享协议的推广，最终实现对不同平台的开放获取数据能高效、快捷的采集。以元数据为基础，借助于大数据、数据挖掘等技术，使开放获取资源更好的为科研人员服务。

参考文献

[1]Budapest Open Access Initiative[EB/OL]. [2016-04-23]. http://www.soros.org/openaccess/read.shtml.

[2]DOAJ[EB/OL].[2016-04-26]. https://doaj.org/

[3]COAJ[EB/OL].[2016-04-26]. http://www.oaj.cas.cn/

[4]王丹. 开放获取仓储发展研究[D].东北师范大学,2011.

[5]黄金霞,陈雪飞,沈东婧. 开放资源的利用评价研究[J]. 图书情报工作,2013,21:18-23.

[6]韩月萍. 网络开放存取的学术资源及其获取策略[J]. 青海师范大学学报(哲学社会科学版),2010,05:163-166.

[7]李武. 开放存取出版的两种主要实现途径[J]. 大学图书馆学报,2005,04:58-63.

[8]王芳,王小丽. 基于OAI协议的数字档案馆元数据互操作问题研究[J]. 现代图书情报技术,2007,03:18-24.

[9]徐方,张静. 国内OAI-PMH协议研究综述[J]. 现代情报,2009,01:89-94.

[10]OAI[EB/OL].[2016-04-26]. http://www.openarchives.org/Register/BrowseSites

[11]王思丽,马建玲,王楠等. 开放知识资源的元数据自动采集策略研究[J]. 图书馆学研究,2013,12:47-51.

[12]陈博. 基于OAI-PMH的数据提供在基础医学信息共享系统中的应用与实现[D].吉林大学,2005.

[13]郭少友. OAI-PMH元数据的关联数据化方法研究[J]. 图书情报工作,2011,02:107-111.

[14]杨文柱,徐林昊,陈少飞等. 基于XPath的Web信息抽取的设计与实现[J]. 计算机工程,2003,16:82-83+113.

[15]蒲筱哥. 基于Web的信息抽取技术研究综述[J]. 现代情报,2007,10:215-219.

[16]连小刚. 基于DOM的Web信息抽取系统设计与实现[D].华中科技大学,2009.

[17]曲杰涛. 基于DOM的智能网页信息抽取技术研究[D].中国海洋大学,2009

[18]李传席. 基于本体的自适应Web信息抽取方法研究[D].中国科学技术大学,2012

[19]祝美莲. 半结构化网页的信息抽取技术研究[D].中国石油大学,2011.

[20]Alberto H. F. Laender,Berthier A. Ribeiro-Neto,Altigran Soares da Silva,Juliana S. Teixeira. A Brief Survey of Web Data Extraction Tools.[J]. SIGMOD Record,2002,31:.

[21]轩艳艳. 基于XML的Web信息抽取研究与实现[D].武汉理工大学,2008.

[22]贺智平. Web信息自动抽取技术研究[D].西安电子科技大学,2006.

[23]李猛. 基于DOM的Web信息抽取技术的研究与实现[D].大连理工大学,2008.

[24]董娟. 基于页面结构分析的网页信息抽取方法研究[D].中国石油大学,2010.

[25]于静. 基于页面主体提取的WEB信息抽取技术研究[D].南京邮电大学,2013.

[26]吴晓彦. 基于结构语义熵的互联网商品信息抽取技术研究[D].复旦大学,2009.

[27]Stephen Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text[J]. Machine Learning,1999,341.

[28]拜战胜. SNAX系统数据采集研究[D].中南大学,2009.

[29]Kushmerick N. Wrapper Induction for Information Extraction [D]. University of Washington, 1997.

[30]杜鹏. 基于视觉特征的WEB页面信息抽取技术的研究[D].西北师范大学,2009.

[31]Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo. RoadRunner-Towards Automatic Data Extraction from Large Web Site[C]. Proceedings of the 27th VLDB Conference, 2001.

[32]Arocena G.O, Mendelzon A O.WebOQL: Restructuring Documents Databases and Webs. Proceedings of the 14th IEEE International Conference on Data Engineering ,Orlando,Florida,1998:24-33.

[33]Deng C.,S.P.Yu,J.R. Wen, et al. VIPS:A Vsion-based Page Segmentation Algorithm [R]. Microsoft Technical Report, MSR-TR-203-79, 2003.

[34]张伟. 基于视觉特征的Web信息抽取技术的研究与实现[D].华东师范大学,2008.

[35]八爪鱼采集器[EB/OL].[2016-04-10]. <http://www.bazhuayu.com/about#about-us>

[36]火车采集器[EB/OL].[2016-04-10]. <http://www.locoy.com/product>

[37]ET采集器[EB/OL].[2016-04-10]. <http://www.zzcity.net/>

[38]网络矿工采集器[EB/OL].[2016-04-10]. http://www.minerspider.com/product/index.aspx

[39]Import.io[EB/OL].[2016-04-10]. https://www.import.io/

通信地址:中国农业科学院农业信息研究所5309科室，北京，100081

联系方式:17888802420

E-mail：591701477@qq.com

1. \*本文系中国农业科学院科技创新工程项目，转基因生物新品种培育重大专项子课题（项目编号：2014ZX0800940B）研究成果之一。 [↑](#footnote-ref-1)