

国际生物医学科学数据审编研究热点探析*

侯丽 吴萌 侯震 李姣

(中国医学科学院医学信息研究所 北京 100020)

[摘要] 以 2015 - 2016 年两届国际生物医学审编会议上领域专家讨论的报告和会议论文集, 以及 PubMedCentral 中近 5 年来有关 Biocuration 和 Data Biocuration 的研究文献为数据源, 采用内容分析法对生物医学科学数据审编的研究主题进行分析、归纳和总结, 重点从 Biocuration 的工作机制、生物医学数据标准构建与应用、集成与可视化、审编与应用、生物医学文本挖掘等方面进行梳理, 以期为我国生物医学科学数据审编的发展提供国际经验。

[关键词] 生物医学数据; 数据审编; 生物医学文本挖掘; 生物医学数据标准与本体

[中图分类号] R - 056 **[文献标识码]** A **[DOI]** 10.3969/j.issn.1673-6036.2017.05.017

Analysis on the Research Hotspots of International Data Biocuration HOU Li, WU Meng, HOU Zhen, LI Jiao, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China

[Abstract] The paper takes the reports and conference proceedings discussed by domain experts during 2015 - 2016 International Biocuration Conference and the research literatures about biocuration and data biocuration in PubMedCentral in recent 5 years as the data sources, analyzes, concludes and summarizes the research subject of biocuration through the content analysis method, and focuses on the sorting of working mechanism of biocuration, construction & application, integration & visualization, review and editing & application of biomedical data standards, mining of biomedical texts, in order to provide international experience for the development of biocuration in China.

[Keywords] Biomedical data; Data biocuration; Biomedical text mining; Standard and ontology of biomedical

[收稿日期] 2017 - 01 - 11

[作者简介] 侯丽, 博士, 副研究员, 发表论文 20 余篇; 通讯作者: 李姣。

[基金项目] 中央级公益性科研院所基本科研业务费资助课题“生物医学领域大规模语义计算关键技术研究”(课题编号: 2016ZX330010); 国家社科基金项目“面向知识服务的公众健康知识组织体系构建研究”(项目编号: 14BTQ032); 国家重点研发计划“精准医学本体和语义网络构建”(课题编号: 2016YFC0901901)。

1 引言

"Curation" 这个概念是在 2001 年由数字保存联盟与英国国家空间中心举办的 "Digital Curation: Digital Archives, Libraries, and E-Science" 国际研讨会中首次提出的^[1], 美国使用 Data Curation, 国内对其中文译称有“数据监管”^[2]、“数据监护”^[3]。Data Curation 强调对数据积极、持续地有效管理, 进而实现数据增值与再利用^[4-6]。美国 2002

年加入到该领域^[7], 国内有关研究自 2011 年前后开始, 以暨南大学图书馆杨鹤林老师发表的“数据监护: 美国高校图书馆的新探索”作为开端。

在生物医学领域, 近年来随着分子生物学和遗传学迅猛发展, 生物医学文献数量飞速增长, 为生命科学研究者提供了大量的信息与知识, 但文献中研究结果的“可信度”良莠不齐, 给基于文献的审编工作带来很大挑战; 当然, 更大的挑战来自于大规模、高通量组学数据及其生物样本注释数据的审编、注释。正是在这样一个背景下, 生物审编 (Biocuration) 为生物医学领域数据的规范管理提供了可行思路。国际生物研究组织给 Biocuration 的定义^[8]是: 对生物相关的各类信息进行翻译、整合并将其整合进数据库系统, 或者整合到科技文献和大数据集合的资源中, 对生物医学知识精准且全面的表示不仅可便利科学工作者获取数据, 而且是大数据分析的基础, 其目标是融合生物审编专家、软件开发工程师和生物信息学领域研究人员的共同努力。

Biocuration 第 1 次国际会议于 2005 年在美国加州举行, 其后每隔一年或两年举办一次, 中国研究者也在 2010 年陆续加入到 Biocuration 的研究领域并于 2015 年成功承办第八届国际生物医学审编会议^[9], 主要围绕从大数据到知识发现展开了交流与讨论。2016 年生物审编国际会议于 2016 年 4 月 9 - 14 日在瑞士日内瓦召开, 会议主要围绕生物医学信息学领域的前沿理论、先进技术、相关成果进行全面展示与交流, 主要包括生物医学数据的集成与可视化、生物医学数据标准本体的构建与应用、组学数据的功能审编、生物医学数据审编的质量控制、标准与存储、生物医学文本挖掘、生物医学数据库建设、生物医学数据审编与应用等主题。

Biocuration 是生物医学领域和信息科学领域的重要研究热点之一, 本文拟采用内容分析法对近年来, 国际生物医学科学数据审编的研究主题进行分析归纳和总结, 试图从 Biocuration 工作流程、生物医学数据标准构建与应用、生物医学数据的集成与可视化、生物医学数据审编的质量控制与标准、生物医学数据审编注释与应用、生物医学文本挖掘等方面厘清 Biocuration 的研究热点, 以期为国内

Biocuration 的研究与发展提供经验。为了保证对 Biocuration 领域热点的全面采集, 本文通过两种渠道确定待分析文献数据集: 一是收集 2015 - 2016 年两届国际生物医学审编会议上领域专家讨论的报告及会议论文; 二是以“[Abstract] = data curation”、“biocuration”、“[MeSH Terms] = data curation”为检索词在 PubMed Central 中进行文献检索 (检索时间限定在近 5 年内), 剔除非相关文献后, 共得到有效相关研究论文 64 篇。

2 生物医学数据审编的关键问题

2.1 Biocuration 的工作流程

为了使得生物医学数据库更具有价值与相关性, 并且保持数据库中数据资源的实时更新, 一项典型的生物医学数据审编应该包含以下过程^[10]: (1) 对数据提交方提交的初始数据和原始数据进行有效管理, 以确保数据的完整性和语法互操作性。(2) 对文献内容进行详尽研究, 抽取与提交数据密切相关的信息。(3) 使用本体和受控词表对数据中的术语及其关系进行正规表达, 确保这些数据与其他数据库的语义互操作。(4) 数据库中术语映射及术语唯一标识符标注, 进而丰富其与已知知识之间的关联。(5) 将信息加载到数据库中, 以便目标用户使用。生物医学数据审编是对生物医学数据进行利用与重利用不可或缺的环节, 已经成为生物医学信息学的一个重要组成部分并在将生物医学数据带入科学大发现中发挥至关重要的作用。

2.2 生物医学数据标准构建与应用

2.2.1 生物医学数据审编标准重要性探讨
Bloom^[11]在第九届生物审编国际会议的大会报告“Publishing Articles and Making Linked Data Available: Steps Towards Seamless Research Reporting”中指出, 目前在许多领域, 研究报告基于非常良好的文档结构和注释数据集, 对于以后重用非常有利; 但在其他一些领域仍存在研究报告和底层数据库连接不良的问题, 其主要原因在于标准不统一并指出公共科学图书馆 (PLOS) 期刊数据策略、RDA 团

队在数据发表中制定数据发表的各种规范,方便数据的重利用。Maglott^[12]参与多项数据审编项目,涉及医学索引、杂种细胞数据库(HDB)、美国标准菌库(ATCC)目录以及美国国家生物技术信息中心(NCBI)数据库资源建设等,通过参与这些数据审编项目的研究工作,发现目前表型、序列变异、序列和表型关系等领域都面临信息标准化的挑战。Haendel^[13]指出,个体生物的健康与基因和环境的相互影响息息相关,但目前缺乏相对科学的标准来表示表型或者是可能导致该表型的环境因素。所以,如何帮助表型数据创建者来对生物学数据审编做出贡献,如何追踪这些贡献以及提供合适的表现型归因,如何利用表型数据来确诊疾病及判断进化的关联性,审编者需要组成一个新的群体致力于解决这些问题。

2.2.2 生物学数据审编标准构建 关于生物学数据标准本体的构建与应用的探讨,重点围绕生物学本体的构建技术展开,通过基于生物学数据标准化的数据形式构建的本体对于后续展开数据的集成和提取有很大帮助。Bairoch等^[14]构建了一套 Cellosaurus 数据库,该数据库包含无限增殖和自然不死细胞系以及有限细胞系,提供以人类、老鼠、兔子为主的脊椎动物细胞系和无脊椎动物细胞系;从数量和信息利用价值上对 Cellosaurus 进行不断扩充,包括添加癌症细胞系对药物的抵抗、化合物、杂种细胞生成的单克隆抗体的靶蛋白以及使一个细胞系不死的转化方法;与国际细胞系认证委员会密切合作来鉴定问题细胞系,使用良好的命名规则。Jia等^[15]提出了一个名为儿科注释与医学(PAM)的标准数据库,该数据库集成了全世界的儿科临床数据来支持临床应用和基础研究,并且基于国际疾病分类法(ICD-10)、国际系统医学术语表——临床术语(SNOMED-CT)和疾病本体(DO)实现了儿科疾病的标准命名与分类。同时,该数据库也提供了一个信息交换平台供使用者录入新知识,在一定程度上实现了疾病和用药的术语标准化和手动审编。Hinard等^[16]介绍了离子通道在不同生物过程中所扮演的不同角色,以及由于离子通道突变所引起的多种疾病。由于在分子水平缺乏离

子通道功能的标准化描述,离子通道电生理学本体(ICEPO)应运而生,ICEPO允许对离子通道电压门控类别的电生理学参数进行注释。

2.3 生物医学数据的集成与可视化

在生物学数据集成与可视化的研究主题中,对异构数据的集成和可视化展示技术与工具的探讨是该主题的研究热点,通过对生物学数据的集成与可视化分析,能够更加直观地揭示基础研究和疾病发现之间更深层次的关系。Wu等^[17]指出,蛋白翻译后修饰 Protein Post-Translational Modification (PTM)是重要的细胞调节机制,而且PTM的破坏与疾病的发生有关联。随着对PTM研究的深入,科学文献中PTM信息越来越多,据此开发了一套网络资源系统^[18],该系统能从数据库和文献挖掘及本体中集成PTM信息并提供搜索服务和序列对比,进一步可开发PTM网络和关联的可视化工具,方便其他研究者对从数据库与文献挖掘出的疾病蛋白关系进行可视化展示。Roux等^[19]指出,基因本体的富集性检验已经成为发现大量基因序列信息的标准,可以通过基因与GO term之间的关联,发现基因序列中高表达或低表达的GO种类,据此提出了一个基因集富集性分析的新应用,基于从基因表达数据计算得来基因和Uberon本体(解剖本体)中生物解剖注释之间的关系,并以可视化工具进行展示。

2.4 生物医学数据审编与应用

2.4.1 组学数据的功能审编 组学数据是生物学数据的一个重要分支,为了让不同数据库之间建立有效的关联,需要对来自不同文献与试验中的组学数据功能进行科学审编。Aimo等^[20]指出,脂肪是种类繁多、数量巨大的生物分子,并且在膜形成、能量储存和信号传递以及对疾病的潜在诊断方面都发挥着特定的作用。现代分析技术如高通量串联质谱分析可以对脂肪组进行分析,但是全面了解脂肪在人类健康中扮演的角色仍需要集成脂肪组数据和生物学知识,因此,Aimo等开发了一套脂肪组及相关生物学知识资源库——SwissLipids。SwissLipids提供层次聚类方法将质谱分析结果和500 000个

脂肪结构进行链接, 基于专家审编的相关数据将代谢途径和主要文献以及证明序列进行集成。总的来说, SwissLipids 为脂肪组数据的发表、数据挖掘和生成假设提供了参考名字空间, 且数据每天更新, 所有数据都可从其官网免费下载。Thomas 等^[21]指出, 基因本体 GO 的注释范例已经主宰功能注释领域很多年, 但是 GO 的注释目前来说太简单, 通常将一个基因产物关联一个单一的功能概念, 所以他们描述了一个新的框架, 将简单的 GO 功能描述与复杂的基因功能描述或者模型连接起来, 并且开发了 Noctua 工具来创建与 GO 注释的链接。

2.4.2 群体审编 Sacristan 等^[22]指出, 目前对于精神病的研究认为精神疾病呈现出很强的遗传因素, 但是对于导致精神病的细胞以及分子机制等的因素了解较少。PsyGeNET 集成了大量的精神疾病遗传学信息, 以标准化的信息格式呈现, 并且提供一套分析和可视化工具。更新 PsyGeNET 的过程包括: (1) 使用文本挖掘技术从文献中挖掘基因与疾病的关系, 并且基于标准的生物学本体和词汇将挖掘出的实体进行标准化。其中, 使用的文本挖掘工具是 BeFree, 该工具在识别基因上可以达到 85% 的 F 分数。(2) 精神科学方面的专家对文本挖掘信息进行人工审编, 开发了一个可以多人使用的网站来作为标注工具。(3) 与其他公开数据库的资源进行整合。Wagmeester 等^[23]指出, 由于建立、维护、更新、集成生物数据库的过程是时间和资源密集型的, 但群体审编与科学知识的产生速率还很难达到平衡, 所以可以利用维基百科的大量已存在的群体来进行群体审编。生物审编群体将维基百科作为一个推广的渠道、一个集成数据的平台和群体审编的集中资源, 其团队已经开始从维基百科知识库中集成药物、疾病、基因等多种数据, 并与审编者进行紧密合作, 接下来希望能找到系统的方法来从维基百科群体中获取知识。

2.4.3 大规模预测注释 Pedruzzi 等^[24]指出, 高质量自动及人工蛋白注释系统 HAMAP 是基于规则的自动蛋白质序列功能注释系统。HAMAP 的蛋白质家族文件和注释规则是经验丰富的审编专家利用 UniProtKB/Swiss - Prot 注释实体的试验数据进行创

建的, HAMAP 也可用于直接注释单个的蛋白质序列或微生物蛋白质组。目前, HAMAP 的应用范围也在不断地向真核生物及病毒蛋白质家族扩展。Pagni^[25]提出, MetaNetX 是一个储存基因组规模的代谢网络及生物通路的数据库, 将多来源的主流资源汇入到一个共同的化合物、反应、蛋白等的名字空间中。MetaNetX.org 的网站提供获取集成数据的渠道, 以及用户导入自己的代谢网络数据、与 MNXref 进行映射并进行操作、比对、分析、模拟等相关工具。Zhang^[26]指出, 随着高通量技术的发展, 产生的生物数据的数量和种类呈指数级增长。目前中国需要收集大量数据以及提供简便获取数据方式的资源中心。Big Data Center 大数据中心, 作为中科院北京基因组所的一个组成部分, 于 2016 年 1 月正式建立, 旨在收集、集成、翻译大量数据。BIGD 的 3 大主要特点是: (1) 构建基因组序列存档。(2) 将中国人口的基因组数据和健康数据进行合并。(3) 集成在中国有经济重要性的物种的组学数据。BIGD 将通过大量组学数据的集成、审编来构建维护生物医学数据库, 并研究分析翻译大数据的方法, 提供免费开放获取数据的渠道, 提供支持世界学术、产业研究的工具。

2.5 生物医学文本挖掘

2.5.1 生物医学文本挖掘系统 Korves 等^[27]指出, 专家审编已难以应对海量文献, 所以研究人工和机器相结合的方式来审编是非常必要的。试验数据基于 DARPA Big Mechanism 项目的文本数据, 首先为了验证吞吐率和准确度, 通过人工获取、机器获取的人工与机器相结合的方式获取文献中的生物学关系。试验结果表明人工审编虽然在鉴定相互作用方面有优势, 但是提供正确的 UniProt 识别时准确率很低。而在人工审编的基础上进行机器审编, 提供正确的 UniProt 识别方面效率较高。为了改善人工审编费时费力的现状, Breuza 等^[28]研发了 TextpressoCentral 文本挖掘系统对文本进行直接注释, 并将这些注释集成于多种审编系统中。该系统提供关键字以及语义规则搜索, 文本注释可以简单地集成于用户定义的审编平台, 可以作为训练集

来进一步提高挖掘质量。为了将 TextpressoCentral 系统集成于目前的审编流水线中, 他们与基因本体联盟密切合作, 将该系统集成于基因本体联盟审编框架中。

2.5.2 基于文献的实体识别 Liu 等^[29]提出了一个基于多核的机器学习方法从 PubMed 文献中提取化学药物与疾病之间的关系 CRDS: 首先, 根据不同大小的文本对象构造不同的核函数, 来反映相关文本不同语义特征; 然后构建多核学习框架来将两个或多个单核函数进行结合; 最后, 将算法应用于 BioCreative V 语料库来验证算法的效率, 并按照国际标准对结果进行评估。结果表明, 该算法具有更高的准确度和效率。Li 等^[30]指出利用生物医学数据进行文本挖掘存在挑战, 但对提升生物医学数据的人工审编具有补充作用, 并且通过文本挖掘技术提升了粗糙数据集中发现 SNPs 的性能, 开展基因组数据分析工作流程的优化, 从而发现某一特定基因新的功能结构。Hoehndorf 等^[31]通过对 500 万医学文摘及文献全文数据的挖掘, 借助 OMIM、OrphaNET、SIDER 等知识组织工具, 识别文献中出现的疾病名、药物靶标、候选基因、通路等实体, 通过共现频次去发现通用疾病的表型, 借助 DO、HPO 进行疾病与表型关系的匹配, 进而对疾病群组可能出现表型同质性进行系统分析, 并据此构建了疾病表型的相似性网络。

3 结语

Biocuration 的研究热点主要是围绕着生物医学数据从审编标注、分析到利用的各个环节中的关键问题展开, 包括生物医学数据标准构建与应用、生物医学数据的集成与可视化、生物医学数据审编的质量控制与标准、生物医学文本挖掘与生物医学数据库建设等主题。生物医学数据标准及其数字化是提升资源共享的重要基础, 生物医学数据的审编是保证数据质量与可用的重要环节, 其中群体审编必将成为审编工作的未来发展趋势, 需要领域的专业审编人员一起探讨, 且对生物审编人员的培训也是保证生物医学数据审编质量的重要方面。基于生物

医学文本进行生物医学实体间关系的挖掘与发现为提升生物医学审编的性能提供了自动化的支持, 基于此构建科学、内容丰富的生物医学数据库则为 Biocuration 的应用与发展提供了应用场景。随着研究的不断深入, 生物医学信息技术的发展还将面临更多的机遇与挑战。生物医学信息领域的学者已经开始拓展视角, 将生物医学数据与其他技术相融合, 重视生物医学数据、医学知识库的信息挖掘与应用, 充分开发利用医疗数据蕴含的信息与价值。作为从事生物医学数据研究的研究机构也应紧跟国际步伐, 加强生物医学数据标准建设、知识库建设、生物医学数据审编质量以及生物医学文本挖掘等方面的研究。

参考文献

- 1 Pothen P. Digital Curation: digital archives, libraries, and e - science [EB/OL]. [2016 - 05 - 20]. <http://www.dpconline.org/events/previous-events/309-digital-curation>.
- 2 谭榕, 亓靖涛. 新时代新角色: 数字保存的监管 [J]. 图书与情报, 2011, (3): 45 - 48.
- 3 杨鹤林. 数据监护: 美国高校图书馆的新探索 [J]. 大学图书馆学报, 2011, (2): 18 - 21, 41.
- 4 University of Illinois Board of Trustees. Specialization in Data Curation [EB/OL]. [2016 - 07 - 20]. http://www.lis.illinois.edu/academics/degrees/specializations/dat_a_curation.
- 5 Lord P, Macdonald A. Data Curation for E - science in the UK: an audit to establish requirements for future curation and provision [EB/OL]. [2016 - 07 - 20]. <https://www.cs.york.ac.uk/ftpdire/pub/leo/york-msc-2007/information/vsr-curation/science-dc-report.pdf>.
- 6 Lord P, Macdonald A, Lyon L, et al. From Data Deluge to Data Curation [EB/OL]. [2016 - 07 - 20]. <http://www.allhands.org.uk/2004/proceedings/papers/150.pdf>.
- 7 Gray J, Szalay A S, Thaka RAR, et al. Online Scientific Data Curation, Publication, and Archiving [EB/OL]. [2016 - 06 - 15]. <http://arxiv.org/abs/cs/0208012>.
- 8 International Society for Biocuration. What is Biocuration [EB/OL]. [2016 - 07 - 12]. <http://biocuration.org/dissemination/who-are-we/>.

- 9 International society Biocuration. International Biocuration Conference [EB/OL]. [2016-07-12]. <http://biocuration.org/community/conferences/international-biocuration-meetings/>
- 10 Zhang Z, Zhu W, Luo J. Bringing Biocuration to China [J]. *Genomics, Proteomics & Bioinformatics*, 2014, 12 (4): 153-155.
- 11 Bloom T. Publishing articles and making linked data available; steps towards seamless research reporting [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 5-12.
- 12 Maglott D. From Paper to the Cloud; the mechanics of biocuration may change but not the logic [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 14-22.
- 13 Haendel M. Why the World Needs Phenopacketeers, and How to be One [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 24-30.
- 14 Bairoch A. The Cellosaurus, a Wealth of Information on Cell Lines [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 164-171.
- 15 Jia J, Shi T. PAM: a standards-based database for integrating and exchanging pediatrics-specified information from multitude of biomedical resources [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 58-66.
- 16 Hinard V, Britan A, Rougier JS, et al. ICEPO: the ion channel electro physiology ontology [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 5-12.
- 17 Wu C, Arighi C, Huang H. iPTMnet: an integrated resource for protein post-translational modification network analysis [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 157-163.
- 18 Protein Information Resource. iPTMnet [EB/OL]. [2016-07-20]. <http://proteininformationresource.org/iPTMnet>
- 19 Roux J, Seppey M, Sanjeev K, et al. TopAnat: GO-like enrichment of anatomical terms mapped to genes by expression patterns [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 104-109.
- 20 Aimo L, Liechti R, Hyka-Nouspikel N, et al. The Swiss-Lipids Knowledge Resource for Lipid Biology [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 75-79.
- 21 Thomas PD, Mungall CJ. LEGO: expressing complex biological models using the Gene Ontology [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 136-144.
- 22 Sacristan AG, Bravo A, Valverde O, et al. Leveraging Text Mining, Expert Curation and Data Integration to Develop a Database on Psychiatric Diseases and Their Genes [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 26-30.
- 23 Wagmeester A, Mitraka E. Leveraging Wikidata for crowd curation [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 37-43.
- 24 Pedruzzi I, Rivoire C, Auchincloss AH. HAMAP - leveraging Swiss-Prot Curation for the Annotation of Uncharacterized Proteins [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 89-96.
- 25 Pagni M. MetaNetX/MNXref - reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 118-125.
- 26 Zhang Z, Big Data Center Members. the BIG Data Center; from deposition to integration to translation [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 151-156.
- 27 Korves T, Garay C. Exploring Human-Machine Cooperative Curation [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 126-135.
- 28 Breuza L, Magrane M, Consortium U. The UniProtKB Guide to the Human Proteome [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 145-150.
- 29 Liu Y, Sun Y, Hou L, et al. Identifying Chemical-disease Relationships Using Multiple Kernel Learning Method [C]. Geneva, Swiss; Proceedings of the 9th International Biocuration Conference, 2016; 185-190.
- 30 Li Y. Challenges and Practices of Big Data in Life Science [C]. Beijing, China; Proceedings of the 8th International Biocuration Conference, 2015; 145-152.
- 31 Hoehndorf R, Schofield P, Gkoutos G. Large-scale Semantic Mining of Disease-phenotype Annotations [C]. Beijing, China; Proceedings of the 8th International Biocuration Conference, 2015; 145-152.